Check for updates

Research Report

Modeling Writing Traits in a Formative Essay Corpus

ETS RR–24-02

Paul Deane
Duanli Yan
Katherine Castellano
Yigal Attali
Michelle Lamar
Mo Zhang
Ian Blood
James V. Bruno
Chen Li
Wenju Cui
Chunyi Ruan
Colleen Appel
Kofi James
Rodolfo Long
Farah Qureshi

December 2024

ETS

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Senior Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Director Research

Larry Davis
Director Research

Jamie Mikeska
Senior Research Scientist

Gautam Puhan
Principal Psychometrician

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Senior Research Scientist

Michael Walker
Distinguished Presidential Appointee

Klaus Zechner
Senior Research Scientist

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Modeling Writing Traits in a Formative Essay Corpus

Paul Deane, Duanli Yan, Katherine Castellano, Yigal Attali, Michelle Lamar, Mo Zhang, Ian Blood, James V. Bruno, Chen Li, Wenju Cui, Chunyi Ruan, Colleen Appel, Kofi James, Rodolfo Long, & Farah Qureshi

ETS, Princeton, NJ

This paper presents a multidimensional model of variation in writing quality, register, and genre in student essays, trained and tested via confirmatory factor analysis of 1.37 million essay submissions to ETS' digital writing service, Criterion®. The model was also validated with several other corpora, which indicated that it provides a reasonable fit for essay data from 4th grade to college. It includes an analysis of the test-retest reliability of each trait, longitudinal trends by trait, both within the school year and from 4th to 12th grades, and analysis of genre differences by trait, using prompts from the Criterion topic library aligned with the major modes of writing (exposition, argumentation, narrative, description, process, comparison and contrast, and cause and effect). It demonstrates that many of the traits are about as reliable as overall e-rater® scores, that the trait model can be used to build models somewhat more closely aligned with human scores than standard e-rater models, and that there are large, significant trait differences by genre, consistent with genre differences in trait patterns described in the larger literature. Some of the traits demonstrated clear trends between successive revisions. Students using Criterion appear to have consistently improved grammar, usage, and spelling after getting Criterion feedback and to have marginally improved essay organization. Many of the traits also demonstrated clear grade level trends. These features indicate that the trait model could be used to support more detailed scoring and reporting for writing assessments and learning tools.

Keywords writing; assessment; natural language processing; NLP; automated essay scoring; AES; automated writing evaluation; AWE; factor analysis; structural equation modeling; SEM; multidimensional model; corpus linguistics; genre; register; writing quality

doi:10.1002/ets2.12377

Background

Written texts vary across multiple dimensions. Some of these dimensions affect writing quality (Diederich et al., 1961) and must therefore be addressed during instruction. Teachers often present sample texts that model such traits (Gallagher, 2011) and train students to use them to evaluate and revise their writing (Culham, 2003). However, analyzing a text in terms of multiple traits can be both laborious and time consuming (Weigle, 2002). This fact suggests that there may be significant instructional advantages to producing trait scores using automated writing evaluation (AWE) systems, because AWE feedback can reduce the burden of scoring for teachers and enable fast feedback loops that supply students with personalized feedback.

For the most part, the features employed in AWE scoring models have straightforward (often linear) relations to writing quality. Essays are better, as a rule, when they are well developed and organized; when they use a richer vocabulary; and when they use language correctly, avoiding errors in grammar, usage, and mechanics. Existing AWE trait analysis systems use such features to define reportable traits, using supervised or unsupervised methods. Supervised methods predict human trait score judgments from AWE features (Shermis et al., 2002). Unsupervised methods identify latent writing traits by analyzing the factor structure of AWE feature sets (Attali & Powers, 2008; Attali & Sinharay, 2015; Crossley & McNamara, 2014; Kim & Crossley, 2018). There is evidence that the resulting systems can effectively support student learning. For instance, feedback systems based on automated trait analysis report significant improvements on specific traits when students revise their essays (Foltz & Rosenstein, 2019).

But not all dimensions of textual variation map directly onto differences in writing quality. Some differences among texts reflect differences among genres (Frow, 2014), as the same feature may occur frequently in one genre but rarely

Corresponding author: P. Deane, E-mail: pdeane@ets.org

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in another (Biber, 1989, 1991; Biber et al., 2002). Some dimensions of text variation also have strong associations with readability (Brinton & Danielson, 1958; Deane et al., 2006; Entin & Klare, 1978) or with patterns of linguistic development (Haswell, 2000; Reppen, 2007).

Of course, differences in genre characteristics and readability can affect text quality. A well-written text is more likely to respect the norms of its genre and be pitched at a reasonable reading level for its audience (Allen et al., 2016). But relations between text characteristics and writing quality may be indirect and involve trade-offs among multiple standards of evaluation. For instance, revising a text to make it more cohesive (and hence easier to read) may cause it to seem wordy and repetitive, at least to readers with high levels of knowledge (McNamara et al., 1996). Writers deal with such trade-offs all the time. Therefore there may be value in creating a writing trait model that is sensitive to other factors, such as those that account for variation among genres or differences in readability, rather than selecting features based solely on their utility as predictors of writing quality.

This research report is designed to explore whether we can build a general writing trait model that can be validated in multiple ways: by modeling text quality, developmental patterns, and genre differences in student writing. Our analysis builds on three major prior lines of research: (a) a series of AWE studies that explored measurement properties of the features deployed in the e-rater® automated scoring engine (Burstein et al., 2004), (b) research into genre variation and readability features that underlie the TextEvaluator® text readability tool for predicting the reading levels of texts (Sheehan et al., 2014), and (c) research into the value of AWE features to predict external measures and provide useful feedback in a formative writing system (Burstein et al., 2018; Burstein et al., 2017).

Feature Sets Used

e-rater

The e-rater automated scoring engine developed at ETS is designed to predict human essay scores from features designed to capture distinct aspects of writing quality: organization and development; discourse coherence; word choice; syntactic variety; idiomatic language; and avoidance of errors in grammar, usage, mechanics, and style. It generally achieves strong agreement with human scores, comparable to the level of agreement found between two different human raters (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013). It can be viewed as measuring most of the constructs in the popular six-trait approach to evaluating student writing (Quinlan et al., 2009), though it does not evaluate voice or the quality of ideas or constructs associated with an extended writing process (Deane, 2013).

E-rater features have been shown to be predictive of a variety of external measures, including class grades, summative end-of-year reading and writing assessments, and admissions tests that aid in making high-stakes decisions (Burstein, McCaffrey, Beigman Klebanov, Ling, & Holtzman, 2019; Deane et al., 2019). Exploratory factor analysis of these features supports a three-factor structure: organization and development, vocabulary, and conventions (Attali, 2011; Attali & Sinharay, 2015). A study that examined essay-writing performance in 4th, 6th, 8th, 10th, and 12th grades also indicated that there is a regular increase in writing performance as measured by e-rater across grade levels (Attali & Powers, 2008).

TextEvaluator

The TextEvaluator text readability tool developed at ETS is designed to predict the grade level and the overall difficulty/readability of texts (Sheehan et al., 2014). It is built on an array of AWE features designed to measure factors that affect readability, including academic language, sentence complexity, concreteness and imageability, word unfamiliarity, referential and lexical cohesion, interactive/conversational style, degree of narrativity, and argumentation. Some of the features included in TextEvaluator models are designed to help distinguish among texts from different genres, particularly literary versus informational texts, to control for genre differences when assessing readability (Sheehan et al., 2007). TextEvaluator models are trained on a corpus of adult, edited texts selected for use in state or national reading assessments and college admissions assessments that aid in making high-stakes decisions; the Stanford Achievement Test, as part of the definition of the Common Core State Standards; or in previous published studies of text readability (Sheehan, 2016). TextEvaluator generates genre classifications of texts (as informational or literary), a grade-level readability estimate with an associated scale score, and a set of factor scores.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Exploratory Natural Language Processing Feature Engine

An exploratory natural language processing (NLP) feature engine was developed as part of an ongoing research project that explored differences among genres and the relationship between AWE features and broader outcomes, such as grade point average and college entrance examinations (Burstein, McCaffrey, Beigman Klebanov, & Ling, 2019; Burstein, McCaffrey, Beigman Klebanov, Ling, & Holtzman, 2019; Burstein et al., 2017). This work included a factor analysis that identified several components that varied across genres and reflected important aspects of the writing construct, including organization and development, coherence, argumentation, and editing (Burstein et al., 2018). This engine underlies the visualizations and feedback provided by the Writing Mentor® online writing practice tool, an AWE system designed to scaffold the revision process for student writers (Burstein et al., 2018; Madnani et al., 2018).

Overlap Between e-rater and TextEvaluator Features

There is relatively little overlap between e-rater and TextEvaluator features. Both systems use word frequency features, but the word frequency measures are based on different text corpora. Both systems use word length features, but TextEvaluator measures word length primarily by using the number of syllables, whereas e-rater uses a transformation of word length in characters. Both systems use measures of text coherence and cohesion, but they are constructed differently. The e-rater discourse coherence feature identifies lexical chains (sequences of repeated or related words) and uses them to build a predictive model intended to maximize association with scores (Burstein, Tetreault, Chodorow et al., 2013). The e-rater style feature is dominated by a measure of word repetition (Quinlan et al., 2009), which is closely related to measures of lexical cohesion. TextEvaluator also starts with lexical overlap (word repetitions) but normalizes these inputs with respect to text length and genre. The resulting features are designed to produce valid comparisons of text cohesion across disparate students and texts (Sheehan, 2013). Otherwise, e-rater and TextEvaluator features appear to measure rather different constructs.

Motivations for Building a Combined Model

There are several reasons to build a trait model for student writing that combines e-rater, TextEvaluator, and exploratory engine features. Perhaps the most important reason is construct coverage. TextEvaluator includes constructs that matter for English language arts instruction yet are not captured by the e-rater engine. Syntactic complexity is an important dimension of student writing development. Concreteness and imageability are closely linked to vividness of language, for they are part of what teachers are interested in when they encourage students to develop a stronger writing voice. Oral or interactive style is important in many genres and social contexts, even if it is typically discouraged in formal written essays. Similarly, the WAVES engine has a number of features that are specifically diagnostic of argumentative and complex academic texts that are not included in the other engines. As students learn to write in different genres, they need to vary how they write to increase or decrease the presence of narrative and argumentative elements.

A second reason is stability of measurement. The features in the e-rater engine are designed for use in linear regression models, which means that collinearity must be minimized. The TextEvaluator and exploratory engines provide additional ways to measure shared constructs, including features that are more sensitive to genre differences. This may yield more stable estimates of underlying student traits. Finally, the enriched feature set may make it possible to build better developmental models of student writing by using features that have already been shown to be predictive on multiple criterion variables.

Role of Essay Length

A potential confounding issue for AWE systems is the typically strong relation between essay score and essay length in words (Deane, 2013). If a writing trait analysis is to support instruction, the factors that describe essay traits should be based on features that capture constructs worth teaching and not on simple measures, such as essay length, that, if emphasized in instruction, might encourage students to adopt inappropriate writing strategies.

However, length may be a prerequisite for traits that are worth tracking. For instance, a well-organized essay must contain content to be organized; by definition, the very concept of text organization applies more naturally to longer than

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to shorter texts. Automated measures of essay organization may be strongly correlated with measures of essay length yet may serve to differentiate between well-organized and poorly organized essays that contain roughly the same content. Similar considerations may apply to other measurable dimensions of text variation. Many of the features that make a text more effective or meaningful have implications for essay length, if only because they require the writer to deploy specific linguistic resources and thus increase word count.

Essay length is also related to fluency. Cognitive models of writing posit that different writing processes compete for working memory, resulting in trade-offs in which low transcription fluency can impede idea generation and self-monitoring and, conversely, in which students working under conditions of high cognitive load may be less fluent in their writing (Kellogg, 2001; McCutchen, 1996). As a result, variations in the length of essays written to the same prompt might reflect the difficulty of the task, the impact of prior knowledge, and working conditions, even for students who know how to produce essays that match expectations for the genre. Therefore length may be related to aspects of writing quality that are hard to measure directly because people with deeper knowledge about a topic—and greater fluency in writing—are more likely to write fluently and produce high-quality texts as a result.

These considerations suggest that the residual role of productivity should be accounted for in a writing trait model. In the existing e-rater automated scoring system, two features capture most of the variance associated with length: the e-rater Organization feature, which measures the log number of discourse text units present in an essay, and the e-rater Development feature, which measures the log length of discourse text units (in words). These features are based on an algorithm intended to identify key essay elements, including introduction, thesis sentence, topic sentences, development segments, and conclusions, and might therefore reasonably be expected to contribute to an overall organization trait. However, it is possible that these features also capture variance related to overall fluency, which is not likely to be captured by individual trait dimensions. Thus, for certain purposes, such as predicting scores or providing feedback, it may be necessary to supplement the trait model by tracking global fluency (as measured by both features combined).

Research Questions

We seek to address the following questions:

1. What traits (dimensions of textual variation) can be measured in student essay writing (Grades 4–12)?
2. How well does the model generalize across populations and tasks?
3. How reliable are these traits?
4. How are the traits related to differences in genre?
5. How are the traits affected by demographic variables?
6. Can the traits be used to support specific educational goals, including (a) measuring growth in specific traits within and across grades, (b) evaluating changes in specific traits after revision, and (c) assessing overall writing quality?

Methodology

Participants

To address all of the research questions, this study examined 1.37 million submissions made to ETS's Criterion® online writing evaluation service by 203,144 K–12 students between 2004 and 2018. These data comprised all scored essay data submitted to Criterion at U.S./North American primary and secondary schools during this time frame.

Criterion is a digital writing product offered by ETS that provides automated scoring and feedback capabilities (Burstein et al., 2004; Ramineni & Deane, 2016). It has been in use since 2004 and is used in a variety of contexts, including primary and secondary schools in the United States, U.S. institutions of higher education, and various institutions around the world, to support instruction of English language learners. Two versions of this service have been offered: the initial version, available between 2004 and 2013, and a major revision released in 2013, primarily affecting the user interface. Criterion uses ETS's automated scoring engine, e-rater, to evaluate student essays and identify potential errors in grammar, usage, mechanics, and style (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013).

No individual demographic data or school-level demographic data were directly collected by ETS. To obtain school-level demographic data, we matched schools as identified in the ETS operational and client relations databases with school-level data from the U.S. Department of Education's National Center for Education Statistics (NCES) public and private

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

school surveys (National Center for Education Statistics, 2003 – 2018a, 2003 – 2018b). In a few cases, districts had included students from multiple schools (typically a high school and an associated junior high or middle school) without recording this fact in their Criterion school hierarchy, and it was necessary to correct the school list by referring to grade-level and class-name information. There were also a few cases for which data were insufficient to match schools to the NCES data or for which a school was absent from the NCES data set for the school years in question. However, we were able to obtain partial or complete school-level demographic data for 893 of the 934 schools in our data set. Most of the schools for which data were missing were private schools absent from the NCES private school survey data for the years in which they used Criterion.

Between 2004 and 2018, students at 689 public schools, 243 private schools, and 2 home schools made submissions to the Criterion system. Almost all of the schools were in the contiguous United States, except for four schools in the U.S. Virgin Islands, two in Hawai'i, and one in Alberta, Canada. Two hundred forty-nine of these schools were in urban communities, 304 were in suburban communities, 122 were in small towns, and 248 were in rural areas. The geographic distribution of schools was as follows:

- Twenty-five percent of the schools were in the Southeast (10 in Arkansas, 1 in Delaware, 1 in the District of Columbia, 21 in Alabama, 81 in Florida, 31 in Georgia, 3 in Kentucky, 14 in Louisiana, 3 in Maryland, 5 in North Carolina, 1 in South Carolina, 43 in Tennessee, and 29 in Virginia).
- Twenty-five percent of the schools were in the Midwest (1 in Indiana, 22 in Illinois, 96 in Indiana, 2 in Kansas, 44 in Michigan, 5 in Minnesota, 12 in Missouri, 1 in Nebraska, 28 in Ohio, 14 in South Dakota, and 8 in Wisconsin).
- Twenty-two percent of the schools were in the Northeast (6 in Connecticut, 5 in Massachusetts, 4 in Maine, 3 in New Hampshire, 52 in New Jersey, 27 in New York, 104 in Pennsylvania, 1 in Rhode Island, and 2 in Vermont).
- Twelve percent of the schools were in the Pacific states (86 in California, 10 in Oregon, and 11 in Washington).
- Nine percent of the schools were in the Southwest (27 in Arizona, 3 in New Mexico, 30 in Texas, and 22 in Oklahoma).
- Three percent of the schools were in the Rocky Mountain states (2 in Colorado, 1 in Idaho, 1 in Montana, 15 in Nevada, and 10 in Wyoming).

The K – 12 schools that used Criterion included 278 high schools, 210 middle schools, and 52 secondary (combined middle/high) schools. There were 211 primary (combined elementary/middle) schools and 88 elementary schools, 28 schools that instructed all grades, and 67 schools with some other combination of grade levels.

The public schools in the Criterion data set were drawn from a total of 401 school districts (including 373 independent school districts, 11 unified school districts, and 13 independent charter districts). Twenty-five percent of these schools reported that fewer than 25% of their students were eligible for free or reduced-price school lunch, 35% reported that between 25% and 50% of their students were eligible for free or reduced-price school lunch, 27% reported that between 50% and 75% of their students were eligible for free or reduced-price school lunch, and 13% reported that more than 75% of their students were eligible for free or reduced-price school lunch. The private schools were predominantly Roman Catholic parochial schools (189 out of 243). The remaining schools represented a broad array of (mostly Protestant) religious orientations, though there were also a small number of Jewish, Islamic, and nonsectarian schools. The NCES public and private school surveys indicated that 30% of the schools that used Criterion were from majority-minority areas. Fourteen percent of Criterion schools reported that more than 25% of their students were Black, 15% reported that more than 25% were Hispanic, 3% reported that more than 25% were Asian, 1% reported that more than 25% were Native American, less than 1% reported that more than 25% were multiracial, and less than 1% reported that more than 25% were Pacific Islanders.

Including students who submitted essays to Criterion across multiple school years, and hence were counted more than once, 3,860 students were classified as being in 4th grade, 8,670 were classified as being in 5th grade, 27,737 were classified as being in 6th grade, 35,518 were classified as being in 7th grade, 39,671 were classified as being in 8th grade, 29,924 were classified as being in 9th grade, 29,158 were classified as being in 10th grade, 21,599 were classified as being in 11th grade, and 14,335 were classified as being in 12th grade. For students from 2013 on, these classifications were derived from information provided by teachers and administrators when classes were created in the Criterion system. For students from earlier years, these classifications were imputed based on the class name, the grade level associated with the assignment, and other information. The details of the imputation method used to create these individual classifications are given in the Procedures section.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To address Research Questions 2 (How well does the model generalize across populations and tasks?) and 6a (Can the traits be used to measure growth in specific traits within and across grades?), we collected data from several sources, including the following:

- A corpus of 65,372 independently collected, human-scored essays was used to train and evaluate e-rater scoring models for Criterion prompts (Burstein et al., 2004). Each essay had been rated at least twice by human raters on a 6-point scale using Criterion's standard holistic rubric. When these ratings differed by more than 1 point, a third score was included in the data set.
- A corpus of 19,914 student essays was collected between 2009 and 2018 as part of prior ETS research studies, mostly in connection with the CBAL™ learning and assessment tool (cf. Bennett, 2011). The students recruited for these studies ranged in grade level from Grade 5 through Grade 9, though the vast majority were middle school students in Grade 7 or 8. The genres that were tested included informational, research, and literary analysis prompts, but the vast majority of the prompts required students to produce argument essays.
- A corpus of 29,907 essays on 10 argument essay prompts was collected in 2018 from candidates taking the HiSET high school equivalency test.
- A corpus of 7,722 essays was collected from candidates who took the GRE® test. Each candidate submitted essays on two tasks (analyze an argument and analyze an issue), representing a total of 42 prompts.
- A corpus of 7,974 human-scored argument essays across 10 prompts was collected from candidates who took the TOEFL® test for entrance into institutions of higher education. The TOEFL data represented two distinct writing prompt types. One prompt type asked the candidate to write an argument essay based on a provided topic. The other asked the candidate to write an essay summarizing and integrating information from an article and a university-level lecture.
- A corpus of 5,976 essays submitted to a TOEFL test preparation massive open online course (MOOC) that ETS administered in 2018 on the EdX platform, using two released TOEFL essay prompts, was collected.
- The Institute for Education Science-funded Writing Achievement Study corpus of 996 first- and second-year college papers (including a range of genres, including essays, literary analyses, and research papers) was collected from three U.S. institutions as part of a federally funded study (Burstein, McCaffrey, Beigman Klebanov, & Ling, 2019; Burstein, McCaffrey, Beigman Klebanov, Ling, & Holtzman, 2019; Burstein et al., 2017) across a range of disciplines.[1]
- The Michigan Corpus of Upper-Level Student Papers (MICUSP), comprising 774 undergraduate college papers, was collected at the University of Michigan (Römer & Swales, 2010) across a range of disciplines.[2]

These data sets differ from the original Criterion corpus in important ways. Most of them represent different populations, and all of them have a different mix of genres. Many of the data sets focus on argument writing. The CBAL and HiSET corpora focus on source-based essay writing. The IES Writing Achievement Study corpus and the MICUSP focus on full-length research papers and other source-based writing tasks.

It is safe to assume that a model built on the Criterion corpus will not fit these other data sets perfectly. Both feature distributions and covariance patterns may differ significantly. However, if the underlying trait model is valid, it should be possible to fit similar models to each corpus and account for discrepancies with the Criterion model, in terms of the specific genre and population biases of each data set. This is the hypothesis that underlies our approach to Research Question 2.

Materials

The writing tasks examined in this study included 436 standard ("topic library") prompts that were provided as part of the Criterion service and approximately 2,989 locally created prompts (reflecting 22,580 distinct class-level assignments) created and assigned by teachers or administrators. These comprised all e-rater-scored, locally created essay tasks assigned in Criterion by K–12 schools between 2014 and 2018.[3]

Procedure

All submissions to the Criterion corpus were made as regular student submissions, responding to assignments created by school staff, using whatever devices students at each school ordinarily used to complete online assignments. ETS's role was restricted to providing training and support to individual schools and school districts when they implemented

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Criterion in their classrooms. The following information was recorded at the time of submission: (a) the text of the submission; (b) a holistic e-rater score based on an automated scoring model associated with the prompt—in some cases, this was a generic model associated with a specific grade level, whereas in other cases, it was a prompt-specific model; (c) errors and advisories intended to identify cases for which the automatically generated score might not be reliable; (d) for submissions after 2013, levels (high/medium/low) on three traits (focused on organization/development, vocabulary, and conventions), representing earlier trait analyses conducted by Yigal Attali and his colleagues (Attali & Powers, 2008); (e) feedback about structural units identified in the essay, such as thesis and topic sentences; (f) a list of error types and locations within the essay, with associated feedback; and finally, (g) a time stamp indicating the time and date of submission. Each individual submission, or attempt, was linked to a specific teacher, student, class, and assignment. If a student made multiple submissions for the same assignment (by default, unless specified otherwise by the teacher for a specific assignment, up to 10 submissions are allowed), a sequential identifier was generated for each attempt. In our analysis, we focus on NLP features calculated from the student text and, secondarily, on the relation of those features and traits derived from them to e-rater scores. Because we do not have teachers' evaluations of the operational Criterion essays, we use the other corpora listed earlier to examine the relation between AWE features and human scores.

Imputation of Individual Student Grade Levels

Grade levels were directly indicated (by class) only for a portion of the data, which were collected since 2013—constituting approximately half of the total. However, multiple sources of information could be used to impute individual grade levels over the entire data set:

- *Name of the class*. A majority of classes were given names that directly or indirectly identified their grade level, such as English I in high school (ninth grade) or 8th Grade Language Arts, Period 1.
- *Grade level associated with the task*. The post-2013 data indicate that class grade levels and task grade levels matched in the vast majority of cases.
- *Nature of the school*. In a senior high school, for instance, all classes can be expected to fall between 9th and 12th grades.
- *Student cohort status*. Students who take the same classes are usually from the same grade (though there are exceptions, particularly in high school).

These regularities were used to create imputed grade levels for each student. First, we created a rule base that used regular expressions to assign classes to grade levels based on class names. Where this information was not available, grade level was imputed by task, based on the grade-level assignment of the task. If a task's grade-level assignment fell above or below the range of grade levels associated with the school, it was corrected to the nearest valid grade. If a student was assigned multiple grade levels across assignments, this was corrected by taking the modal grade level for the student, and if that did not resolve the conflict, the modal grade level for the class. Where class grade-level assignments were available (in the post-2013 data), the resulting statistic agreed exactly with class grade levels 88% of the time. In 96% of the cases, the imputed grade was no more than one grade level of f We therefore used class grade levels, where available, and imputed grade levels otherwise. Any remaining within-student disagreements based on assigned class grade levels were resolved by taking the average of a student's grade-level assignments and rounding up.

Definition of Variables for Month in School Year and Revision Sequence

All essays were associated with a time stamp for time of submission and an assignment ID that made it possible to match multiple submissions for the same assignment. We used these data to map individual "attempts," or submissions, onto variables that identified the month in school year (1–9) and submission sequence (between 1 and 10). There were very few submissions during the summer (June, July, and August), so for purposes of analysis, we mapped submissions from August onto Month 1 (otherwise, September) and submissions from June and July onto Month 9 (otherwise, May).

Postprocessing

When the K–12 Criterion data were extracted, additional processing was performed to associate each essay with modern AWE features. These features comprised the features used to predict essay scores in e-rater Version 19.1.1, the features

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

used to predict genre status, the reading difficulty in TextEvaluator, and selected exploratory engine features that help to distinguish argument essays from narratives and expository essays.

Preparation of Adjudicated Scores for the Human-Scored Criterion Corpus

For the 65,372 human-scored Criterion essays and the CBAL human-scored data set, final adjudicated scores were calculated as follows: When the first and second raters' scores differed by no more than 1 point, the average of the first and second raters' scores was assigned as the final score. When they differed by 2 or more points, the adjudicated score was averaged with either the first or second rater's score, whichever differed less from the adjudicated score. The result was an adjudicated score on a 6-point scale that included half-point intervals.

Analyses Conducted

Research Question 1

To address Research Question 1 (What dimensions of variation can be measured in student essay writing, Grades 4 – 12?), we conducted confirmatory factor analysis using the e-rater, TextEvaluator, and exploratory engine features identified in Table 1.

Table 1 excludes one TextEvaluator feature related to text cohesion that may also have value for our analysis: the lexical tightness score, which is a measure of the extent to which the words in the document frequently co-occur and are therefore likely to be addressing similar topics. In general, the lexical tightness feature is only weakly correlated with the cohesion features and even more weakly correlated with the features associated with other dimensions. Therefore, rather than including lexical tightness in the factor analysis, we chose to treat it as an independent dimension from the beginning, primarily so that we could use it, for instance, in regressions to predict essay score, where it has a clear added value.

In addition, the factor analyses conducted by Biber and his colleagues (Biber, 1989, 1991; Biber et al., 2002) suggest that there will be a strong relationship between the academic language (formality) dimension and two other dimensions: vocabulary length and vocabulary difficulty. We expect highly academic texts to deploy harder, rarer words and to avoid obviously colloquial language. We also expect a strong relationship between the academic language and organization dimensions, because it is intended precisely to measure the kinds of discourse structures characteristic of a formal, written essay. Similarly, there are natural relationships among the contextualization, cohesion, and dialogue features. Narrative texts are likely to be high in contextualization and dialog features and low in cohesion features. Certain other traits also have obvious relationships: syntactically complex structures are likely to appear in longer sentences with fewer errors in grammar and usage, and highly probable word sequences are likely to be correctly spelled. These additional relationships suggest a hierarchical factor structure, as shown in Figure 1.

We tested this theoretical model (17 factors, i.e., writing traits) using confirmatory factor analysis, setting superordinate dimensions as having zero covariance with subordinate dimensions and loading on the features associated with their subordinates. Specifically, we evaluated the fit of the theoretical model using standard metrics (comparative fit index [CFI]; root mean square error of approximation [RMSEA]) and compared the fit for this model with the fit for other, plausible models with simpler structures into which we merged highly correlated factors.

Research Question 2

To address Research Question 2 (How well does the model generalize across populations and tasks?), we wanted to determine whether the model structure and factor loadings we obtained on the training set were valid across a wide range of cases. To answer this question, we adopted the following procedure:

1. We trained the model on half of the Criterion data, which yielded population norms (which could be used to standardize individual features) and factor loadings (which could be used to define a standard, "fixed" model).
2. We applied the fixed model to the second (held-out) half of the Criterion data. Specifically, we normalized individual features using the means and standard deviations from the training set and set factor loadings equal to the factor loadings obtained from the training set. However, the model structure also included factor covariances. We left these covariance terms free (except for those fixed to zero in the original model, which were once again fixed to

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1 Features Used in the Writing Trait Model

Hypothesized dimension	Feature	Feature description
Formality: general (academic language)	Nominalizations	Proportion of nominalizations formed with suffixes like *-ion* and *-al* (types collapsed: variant forms of a word treated together)
	Coxhead academic words	Proportion of words in the document that are in the Coxhead academic word list
	Abstract nouns	Proportion of word types in the document on a list of common abstract nouns
	Cognitive process perception nouns	Words for internal mental states (*thought*, *feeling*)
	Syntactic variety	Based on an algorithm that weights the grammatical elements present in a text by their relative distribution in a well-structured essay (also loads on organization)
Formality: vocabulary length	Average word syllables	Total syllables in a document divided by the number of words
	Long words	Percentage of the words in a document that are longer than eight characters
Formality: vocabulary difficulty	(In)frequency	$-1 \times$ the median word frequency in the Google Books corpus
	Mean TASA SFI	Mean SFI in the TASA corpus
	Mean ETS word frequency	Mean log word frequency in an ETS-compiled corpus
Formality: organization	Essay elements (e-rater Organization feature)	Based on an algorithm that divides the essay into major text units (introduction, thesis sentence, topic sentences, developing material, conclusion)
	Number of lexical chains	Based on an algorithm that identifies lexical chains (sequences of repeated or related words) bounded by a transition word
	Length of lexical chains	Based on an algorithm that identifies lexical chains (sequences of repeated or related words) bounded by a transition word
	Syntactic variety	Based on an algorithm that weights the grammatical elements present in a text by their relative distribution in a well-structured essay (also loads on academic language)
Sentence structure: sentence length	Avg. clause length	Number of words per clause
	Avg. sentence length	Number of words per sentence
	Avg. number dependent clauses	Number of subordinate clauses per sentence
Sentence structure: sentence complexity	Yngve's depth	Average depth of syntactic embedding
	Theme complexity	Average number of words before the main verb
	Avg. sentence length	Cross-loads on sentence length
Sentence structure: grammar and usage	Grammar	Avoidance of grammar errors (negative square root of proportion of grammar errors)
	Usage	Avoidance of usage errors (negative square root of proportion of usage errors)
Conventionality (language and mechanics)	Mechanics	Avoidance of spelling and punctuation errors (negative square root of proportion of spelling and punctuation errors)
	Grammaticality	Natural phrasing (measured by the probability of word sequences—not grammaticality in the prescriptive grammar sense)
	Idiomaticity	Correct usage of collocations and prepositions
Narrativity: contextualization	Past tense verbs	Typical markers of narrative
	Past perfect verbs	Sentences built around grammatical structures typically used in arguments, such as complement clauses and infinitives
	Third person pronouns	Typical markers of narrative
	Fiction verbs	Verbs that typically appear in narratives or conversations, such as *sit*, *walk*, or *look* (cross-loads on interactivity)

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1 Continued

Hypothesized dimension	Feature	Feature description
Narrativity: dialog	Narrative communication verbs	Verbs typically used to express indirect speech (such as *say*, *ask*, or *tell*) (also loads on the interactivity trait)
	Words inside quotes	Proportion of words in text enclosed in quotation marks
Narrativity: cohesion	Informational cohesion	Normalized vocabulary overlap between adjacent sentences in the text, based on informational text patterns
	Literary cohesion	Normalized vocabulary overlap between adjacent sentences in the text, based on literary text patterns
Stance taking	Stance markers	Words like *however*, *seems likely*, *alleged*
	Argument verbs	Words like *believe*, *argue*, *claim*, *rebut*
Interactivity	Conversation verbs	Verbs used to mark direct speech (e.g., *ask*, *say*)
	First person pronouns	*I*, *me*, *my*, *mine*, *we*, *us*, *ours*, etc.
	Contractions	Shortened combinations like *I'm*, *we're*, *you'll*
	Fiction verbs	Verbs that typically appear in narratives or conversations (cross-loads on contextualization)
	(In)frequency	−1 × the median word frequency in the Google Books corpus (cross-loads on vocabulary frequency)
	Theme complexity	Average number of words before the main verb (cross-loads on sentence complexity)
Concreteness	Concreteness	Proportion of words in the document rated as highly concrete
	Imageability	Proportion of words in the document rated as easy to visualize

Note. SFI = standardized frequency index. TASA = Touchstone Applied Scientific Associates standardized word frequency index.

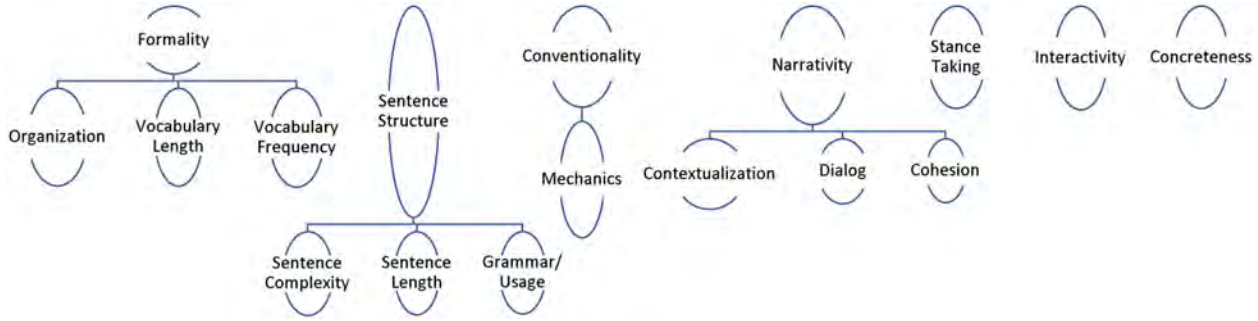

Figure 1 Structure of the proposed writing trait model.

zero) and fitted the resulting model structure to the held-out data so they were normalized. If this model fit well, it was safe to conclude that the overall model structure, the feature normalizations, and the factor loadings obtained from the training set were appropriate for the testing set.

3. We retrained the original model structure (with all factor loading parameters left unset) on the second (held-out) half of the Criterion data. Specifically, we normalized individual features using the means and standard deviations from the test set; left all model parameters free, except for the covariances that needed to be set to zero; and fitted the model to the test set so that it was normalized. If this model (the "self" model) fit well, it was safe to conclude that the underlying structure of the model fit the test data well, even if the feature distributions and factor loadings were meaningfully different from the original training set.
4. We calculated factor scores on the test set for both the "fixed" and the "self" models and then calculated the correlations between corresponding factors. If these correlations were high (above approximately .9), then it was safe to conclude that the differences between the "fixed" and "self" models mostly had to do with differences in feature distributions, not factor loadings.
5. We directly compared the factor loadings on the "fixed" and "self" models to identify sources of misalignment between factor scores in the two models.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We then repeated Steps 2–5 on each of the other data sets: the human-scored essays written to Criterion prompts, the CBAL essays, the HiSET essays, the TOEFL MOOC and TOEFL essays, the GRE essays, the IES Writing Achievement Study essays, and the MICUSP essays. The result of these comparisons should give us an estimate of the extent to which the original, "fixed" model can be applied usefully to data generated by different populations, from a different mix of tasks, or with specific genre or register characteristics.

Research Question 3

To address Research Question 3 (How reliable are these traits?), we examined the test–retest reliability of the traits under conditions very likely to minimize the effect of irrelevant sources of variance:

- The more time that passes from one writing task to the next, the more likely it is that student skill profiles will change. Thus it is better to examine essays written within a short time window, rather than essays written at any time in the school year.
- Student skill profiles are also likely to differ if writing tasks differ significantly in their properties. Thus it may be useful to examine test–retest reliability for essays that were written in the same genre (two persuasive essays or two narratives), rather than comparing trait features across genres. In the Criterion database, genre information is reliable only for assignments based on the Criterion topic library prompts, so it may also be important to restrict the comparison to the topic library prompts.
- Longer essays may provide more evidence from which to infer trait profiles than shorter essays. Thus it may be useful to compare the reliability of longer versus shorter essays.

Our base case, therefore, involves pairs of (final draft) essays written by the same student within the same genre in the same calendar month. This gives an estimate of true test–retest reliability under classroom conditions. If we manipulate variables likely to affect reliability—for instance, by looking only at shorter essays, rather than longer essays, or by extending the time window within which essay pairs were written—then reliability ought to decrease. The reliability of individual traits can be usefully compared to the reliability of overall e-rater scores and to the extent to which trait scores can be predicted from overall (holistic) essay scores. If trait scores are about as reliable as overall scores and not too strongly correlated with overall scores, there may be potential to treat them like subscores and to use them to build a richer profile of student performance than could be obtained from the trait scores alone.

To get a sense of the validity of the trait measures as subscores, we conceptualized e-rater scores as a proxy for total score and regressed the e-rater score on the second essay against trait scores plus the e-rater score on the first essay, using the same database of essay pairs from the Criterion topic library prompts. To the extent that the trait scores are significant predictors of the e-rater score on the second essay above and beyond the e-rater score on the first essay, we can justify use of e-rater scores as subscores (Haberman, 2008).

Of course, ideally, we would evaluate the contribution of the trait scores to test/retest the prediction against human scores, not e-rater scores. However, most of our data sets do not support this comparison. We do, however, have one data set from a pair of CBAL studies in which students wrote two essays in the same genre (randomly paired from a choice of three topics in two genres) within a 3-week period (van Rijn et al., 2016; van Rijn & Yuen, 2016). This data set enables us to create a comparison in which we have 1,669 essay pairs, with trait scores and a human score for each member of the pair. We set up a regression in which each essay plus its human score was used to predict human scores on the other essay in the pair. Because both essays were written within 1 month in the same genre, they represent essentially the same comparison, but using human scores rather than e-rater scores to represent the total holistic score.

Research Question 4

To address Research Question 4 (How are the traits related to differences in genre?), we conducted one-way analyses of variance (ANOVAs) to identify significant trait differences by genre, then performed pairwise Tukey honestly significant difference post hoc tests to determine which trait differences across genres were significant. We then calculated Cohen's *d* and reported effect sizes for significant trait score differences between genres. Moderate or large effect sizes would indicate particularly important differences between genres on specific traits.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Predictive Features Used in e-rater (Excluding Some That Are Used Only in Certain Specialized Models)

Feature label	Feature	Feature description
LOGDTU	Essay elements (e-rater Organization feature)	Based on an algorithm that divides the essay into major text units (introduction, thesis sentence, topic sentences, developing material, conclusion)
LOGDTA	Length of essay elements	Based on an algorithm that divides the essay into major text units (introduction, thesis sentence, topic sentences, developing material, conclusion)
SVF	Syntactic variety	Based on an algorithm that weights the grammatical elements present in a text by their relative distribution in a well-structured essay
DIS_COH	Discourse coherence	Derived from the data underlying the lexical chain and lexical chain length features using a formula that factors out the effect of essay length
NWF_MEDIAN	(In)frequency	$-1 \times$ median word frequency
WORDLN_2	Word character length	Avg. square root of the number of characters in a word
NSQG	Grammar	Avoidance of grammar errors (negative square root of proportion of grammar errors)
NSQU	Usage	Avoidance of usage errors (negative square root of proportion of usage errors)
NSQM	Mechanics	Avoidance of spelling and punctuation errors (negative square root of proportion of spelling and punctuation errors)
NSQS	Style	Avoidance of stylistic issues (negative square root of proportion of stylistic issues)
COLPREP	Idiomaticity	Based on rate of correct usage of collocations and prepositions
GRAMMATICALITY	Grammaticality	Natural phrasing (measured by the probability of word sequences)[a]

[a] Used in some e-rater models.

Research Question 5

To evaluate Research Question 5 (How are the traits affected by demographic variables?), we regressed the available demographic variables against trait scores and estimated effect sizes by adding those regressions to the original structural equation model and refitting the model.

Research Question 6a

To address Research Question 6a (Can the traits be used to measure growth in specific traits within and across grades?), we (a) calculated the correlation between grade levels and trait scores; (b) calculated mean trait scores by genre, grade, and month in school year; and (c) graphed trends in trait scores by grade level and genre (across grades) and by month in school year and genre (within grade). To the extent that there are significant and at least small to moderate grade-level and school year trends, we can conclude that observable growth has occurred, which supports the conclusion that we can use the trait model to measure growth.

Research Question 6b

To address Research Question 6b (Can the traits be used to measure growth in specific traits after revision?), we calculated mean trait scores for successive revisions within groups of students who completed the same number of revisions. This allowed us to observe the extent to which student performance changed after revision after accounting for the fact that students who completed more revisions tended to start with lower-quality essays.

Research Question 6c

To address Research Question 6c (Can the traits be used to assess overall writing quality?), we compared the predictive power of models based purely on e-rater features to the predictive power of models based on the trait model.

E-rater models are built using all or some subset of the features listed in Table 2. Note that most (but not all) of these features are also used in the trait model. Three features were excluded from the trait model: (a) The DIS_COH feature

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(discourse coherence) was excluded from the trait model because the lexical chains and lexical chain length features were more direct measures of essay organization; (b) the WORDLN_2 (word length) feature was excluded because it is mathematically related to the long words feature, which militated against including both in the model; and (c) the NSQS (style) feature was excluded from the trait model because, like the lexical tightness features from TextEvaluator, it generally shows no strong correlations with any of the other features selected for the model.

To evaluate the predictive performance of e-rater features on the data sets for which we have human scores, we split the data 80/20 into training and evaluation sets for the purposes of this analysis. We built two models based on e-rater features. In the first model (Biber, 1989, 1991; Biber et al., 2002), we regressed the 10 regular e-rater features against scores on the training set. In the second model, we regressed all 11 e-rater features (including grammaticality), plus the lexical tightness feature, against scores on the training set. Finally, we calculated the performance of the resulting models on the test set, using the quadratic weighted kappa statistic.

To evaluate the predictive performance of the writing trait model on the data sets for which we had human scores, we used the same training and evaluation sets. Our candidate feature set included the 17 trait scores, the lexical tightness features, and the e-rater LOGDTU and LOGDTA features (to account for the residual effect of composition fluency; see Table 2). We regressed this feature set against scores on the training set. Finally, we calculated the performance of the resulting models on the evaluation set, using the quadratic weighted kappa statistic.

We have essay scores for the Criterion human-scored essay data set, the CBAL essay data set, the HiSET essay data set, the GRE essay data set, and the TOEFL essay data set. We were therefore able to evaluate the performance of the trait model to predict scores across a range of task types and populations.

Results and Discussion by Research Question

Research Question 1: Confirmatory Factor Analysis

Research Question 1: Results

We divided the Criterion corpus randomly (50/50) into training and evaluation sets and trained the model on the training set. When we did so, the 17-factor model demonstrated an acceptable fit (CFI = .934; RMSEA = .0547), falling only marginally below the theoretical threshold for goodness of fit, as advocated by Hu and Bentler (1998; i.e., CFI > .95; RMSEA < .06). When we applied the resulting factor weights to the evaluation set (the "fixed" model), the model fit performed about as well (CFI = .934; RMSEA = .0526). Retraining the model on the evaluation set (the "self" model) yielded essentially the same fit (CFI = .934; RMSEA = .0550). Correlations between trait scores for the fixed and self models were consistently greater than .99.

Table 3 shows the correlations among traits in the model fitted to the training set. The pattern of correlations is more or less what we would predict from the hierarchical structure posited by the model.

We considered four alternative structures: one in which we dropped the mechanics subtrait, one in which we merged the vocabulary length subtrait, one in which we dropped the sentence length subtrait, and one in which we dropped the superordinate formality and narrativity traits. All four showed worse fit than the full model. When we dropped the mechanics subtrait, CFI decreased to .920 and RMSEA increased to .0596. When we dropped the vocabulary length subtrait, CFI decreased to .924 and RMSEA increased to .0580. When we dropped the sentence length subtrait, CFI decreased to .924 and RMSEA decreased to .0580. When we dropped the superordinate factors, CFI decreased to .831 and RMSEA decreased to .0847. Table 4 shows the factor loadings for the full 17-factor model.

Research Question 1: Discussion

Overall, these are very promising results. The 17-factor model fits the Criterion data acceptably well, with features that cohere in theoretically reasonable ways. Each factor has a straightforward interpretation:

- *Formality* is a measure of written, academic style, characterized by longer, rarer, more Latinate vocabulary and an avoidance of markers of oral language, such as first person pronouns, verbs of speaking, and contractions.
    - *Organization* is a measure of how well the essay has been elaborated in well-structured ways, using thesis and topic sentences, transition words, and chains of related ideas, varying the sentence structure to make the relationships between ideas clear.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3 Correlations Between Traits in the Writing Trait Model

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Formality	—																
2. Narrativity	−.11	—															
3. Sentence structure	.06	−.02	—														
4. Sentence length	−.10	.00	**.92**	—													
5. Sentence complexity	.24	−.01	**.83**	**.62**	—												
6. Cohesion	.02	−.11	.04	.03	.00	—											
7. Dialogue	−.03	**.75**	.01	.03	.03	−.10	—										
8. Organization	**.53**	.24	.08	.00	.15	−.03	.22	—									
9. Grammar usage	.13	.00	**.35**	.14	.25	.02	−.02	.10	—								
10. Concreteness	−.17	.22	−.11	−.13	−.03	−.06	.08	.07	−.07	—							
11. Conventionality	.03	−.05	−.08	−.13	−.11	−.02	−.05	.13	.27	−.08	—						
12. Mechanics	−.07	−.04	−.16	−.17	−.21	−.05	−.03	.10	.20	−.08	**.96**	—					
13. Interactivity	**−.63**	.23	−.06	.09	−.26	−.14	.17	−.05	−.03	.07	.14	.24	—				
14. Contextualization	−.14	**.89**	−.03	−.01	−.03	−.21	**.40**	.18	.01	.25	−.05	−.03	.23	—			
15. Stance taking	−.03	**−.35**	.15	.21	−.03	.12	−.10	−.03	.07	**−.31**	.23	.25	.13	**−.42**	—		
16. Vocabulary length	**.88**	−.20	.02	−.13	.19	.05	−.11	.24	.14	−.22	.03	−.06	**−.64**	−.21	−.03	—	
17. Vocabulary frequency	**−.69**	.10	.07	.20	−.14	−.02	.09	−.13	.00	−.03	.17	.26	**.73**	.09	.23	**−.64**	—

Note. Correlations > |.30| are indicated in boldface.

  - *Vocabulary length* and *vocabulary difficulty* represent two other important aspects of formality (word length and word frequency) that give a sense of whether writers have deployed relatively difficult, demanding vocabulary.
- *Sentence structure* represents all aspects of sentence construction, including sentence length, sentence complexity, and grammar/usage.
  - *Sentence length* and *sentence complexity* represent aspects of sentence construction—whether the sentences tend to continue at length and whether they keep sentence structure relatively simple or make use of relatively complex, relatively difficult constructions.
  - *Grammar and usage* represents another aspect of sentence construction—adherence to the grammatical norms for standard written English.
- *Conventionality* and *mechanics* represent the accurate production of normal English words and phrases.
- *Narrativity* represents the group of traits and features that differentiate narratives (on one end) from argument essays (on the other).
  - *Contextualization* represents aspects of sentence structure that people typically deploy in narrative (use of past tense, past perfect, past tense, and fiction verbs).
  - *Dialog* represents the use of indirect speech patterns with the use of narrative communication verbs and quotes.
  - *Cohesion* represents the repetition of key ideas across sentences and is far more prevalent in informational than in narrative texts.
- *Stance taking* represents aspects of vocabulary and sentence structure that people typically use when they are taking a subjective stance, as when they develop an argument.
- *Interactivity* represents typically oral language, marked by contractions, verbs of speaking, and first person pronouns; highly frequent vocabulary typically used in conversations or narratives; and simpler sentence structures characterized by short themes or sentence starts (reflecting the fact that in an oral context, thematic information is highly likely to be given early in the sentence; therefore that information can be subsequently referenced with pronouns or simple noun phrases).
- *Concreteness* represents the extent to which reference is made to physical, concrete, and easily visualized concepts versus abstract ideas.

We add the *lexical tightness* feature (a measure of the extent to which the words in the document frequently co-occur and are therefore likely to be addressing similar topics) to this list as an additional dimension. The net result is a rich multidimensional model that covers many different aspects of student essays.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4 Factor Loadings for the 17-Factor Model

Factor	Feature	Loading
Formality	Average word syllables	1.00
	Long words	.89
	Nominalizations (type collapsed)	.94
	Coxhead academic words	.84
	Abstract nouns (type collapsed)	.64
	Cognitive process perception nouns	.41
	Syntactic variety	.77
	Mean TASA SFI	−.31
	Mean ETS word frequency	−.81
	(In)frequency	.37
	Length of lexical chains	.30
	Number of lexical chains	.43
	Essay elements	.25
Formality: organization	Length of lexical chains	1.00
	Number of lexical chains	.92
	Number of essay elements	.60
	Syntactic variety	.48
Formality: vocabulary length	Average word Syllables	1.00
	Long words	.49
Formality: vocabulary frequency	Mean TASA SFI	1.00
	Mean ETS word frequency	.93
	(In)frequency	−.43
Interactivity	First person pronouns	1.00
	Conversation verbs	.88
	Contractions	.46
	(In)frequency	−.86
	Average word count before main verb	−.46
	Fiction verbs	.35
Concreteness	Proportion highly concrete	1.00
	Proportion highly imageable	.67
Narrativity	Past tense verbs	1.00
	Third person pronouns	.79
	Narrative communication verbs	.87
	Past perfect verbs	.54
	Fiction verbs	.41
	Words inside quotes	.72
	Literary cohesion	.25
Narrativity: contextualization	Past tense verbs	1.00
	Third person pronouns	.79
	Past perfect verbs	.52
	Fiction verbs	.41
Narrativity: dialog	Words inside quotes	1.00
	Narrative communication verbs	.79
Cohesion	Informational cohesion	1.00
	Literary cohesion	.95
Stance taking	Claims	1.00
	Argument verbs	.39
Sentence structure	Mean sentence length	1.00
	Mean clause length	.86
	Mean number of dependent clauses	.67
	Average Yngve's depth	.58
	Average word count before main verb	.32
	Grammar	.62
Sentence structure: sentence complexity	Average Yngve's depth	1.00
	Mean sentence length	.40
	Average word count before main verb	.36
Sentence structure: sentence length	Mean number of dependent clauses	1.00
	Mean clause length	.90
	Mean sentence length	.57
Conventionality	Grammaticality	1.00
	Mechanics	.74
	Idiomaticity of language	.31
Conventionality: mechanics	Grammaticality	1.00
	Mechanics	.22
Conventionality: grammar and usage	Grammar	1.00
	Usage	.84

Note. SFI = standardized frequency index. TASA = Touchstone Applied Scientific Associates standardized word frequencies.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5 Correlations Between Trait Scores for the Fixed and Self Models on Human-Scored Criterion Data

Trait	Correlation
Formality	.95
Vocabulary length	.95
Vocabulary frequency	.92
Organization	.93
Sentence structure	.93
Sentence length	.93
Sentence complexity	.93
Grammar and usage	.93
Conventionality	.94
Mechanics	.94
Narrativity	.91
Contextualization	.91
Dialog	.90
Cohesion	.90
Stance taking	.93
Interactivity	.92
Concreteness	.92

Research Question 2: Model Generalization

Research Question 2: Results

Model Performance on the Corpus of Human-Scored Essays Written to Criterion Prompts

When we applied the fixed Criterion model (i.e., the 17-factor model trained on the original training set) to the human-scored Criterion corpus, the fit was slightly worse than for the operational Criterion corpus (CFI = .916; RMSEA = .0592). When we retrained the model on the human-scored data (producing what we term the "self" model), the fit was slightly better than for classroom Criterion essays (CFI = .941; RMSEA = .0520). Correlations between corresponding trait scores were consistently approximately or greater than .90 (see Table 5).

Model Performance on the CBAL Essay Corpus

When we applied the fixed Criterion model to the CBAL essay corpus, the fit was much lower than what we observed for the classroom Criterion essays (CFI = .805; RMSEA = .0801). However, when we retrained the model on the CBAL data, the fit was still acceptable, though lower than it was for the original Criterion training set (CFI = .901; RMSEA = .0597). Correlations between corresponding trait scores were consistently greater than .90, except for the sentence structure–related traits, which ranged from .85 to .89 (see Table 6).

Upon examining the factor loadings for the least-correlated factors (sentence structure, sentence length, and grammar and usage), we observed that the differences were apparent in the relative magnitude of feature loadings. The CBAL sentence length factor had heavier loadings for mean clause length (ZScore1Mean) and mean sentence length (Avg_Sent_Word_Count; see Figure 2). The CBAL grammar and usage factor had heavier loadings on the grammar feature (NSQG) but weaker loadings on the usage factor (NSQU; see Figure 3). For the superordinate sentence structure factor, the sentence length features had much weaker loadings compared to their loadings in the Criterion data, especially with respect to the features mean clause length and average number of dependent clauses (see Figure 4).

Model Performance on the HiSET Essay Corpus

When we applied the fixed Criterion model to the HiSET essay corpus, fit was lower than what we observed for the operational or human-scored Criterion corpora (CFI = .885; RMSEA = .0649). However, when we retrained the model on the CBAL data, the fit was acceptable and only slightly lower than it was on the operational Criterion training set (CFI = .924; RMSEA = .0553). Correlations between corresponding trait scores were consistently greater than .96 and mostly fell above .99 (see Table 7).

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 6 Correlations Between Trait Scores for the Fixed and Self Models on CBAL Essay Data

Trait	Correlation
Formality	.96
Vocabulary length	.97
Vocabulary frequency	.99
Organization	1.00
Sentence structure	.85
Sentence length	.89
Sentence complexity	.98
Grammar and usage	.87
Conventionality	.99
Mechanics	.98
Narrativity	.92
Contextualization	.98
Dialog	1.00
Cohesion	1.00
Stance taking	1.00
Interactivity	.96
Concreteness	.97

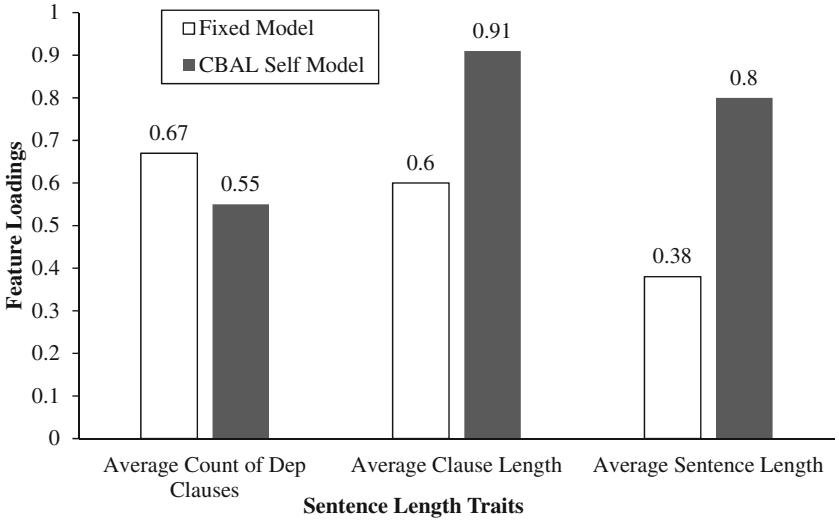

Figure 2 Relative loadings on the sentence length trait for the fixed model versus for a model trained on the CBAL essay corpus.

Model Performance on the TOEFL Massive Online Open Course Corpus

When we applied the fixed Criterion model to the TOEFL MOOC essay corpus, the fit was much lower than what we observed for the operational Criterion corpus (CFI = .822; RMSEA = .0825). However, when we retrained the model on the TOEFL MOOC data, the fit was acceptable and only slightly lower than it was on the original Criterion training set (CFI = .925; RMSEA = .0538). Except for narrativity, $r = .66$, sentence structure, $r = .89$, and contextualization traits, $r = .67$, correlations between corresponding trait scores were consistently greater than .90 (see Table 8).

When we examined the factor loadings for interactivity, we found that weight for fiction verbs was negative, rather than positive, as in the fixed model (see Figure 5). When we examined the factor loadings for contextualization, we found that weight for third person pronouns was negative, rather than positive, as was observed in the fixed model. In addition, the weights for past tense and past perfect verbs were much reduced compared to the fixed model (see Figure 6). When we examined the factor loadings for narrativity, we found that the weight for third person pronouns was negative, rather than positive, as in the fixed model, whereas the weight for the two dialog features (narrative communication verbs and words inside quotes) were almost zero. However, the weight for fiction verbs was much increased compared to the fixed model (see Figure 7).

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

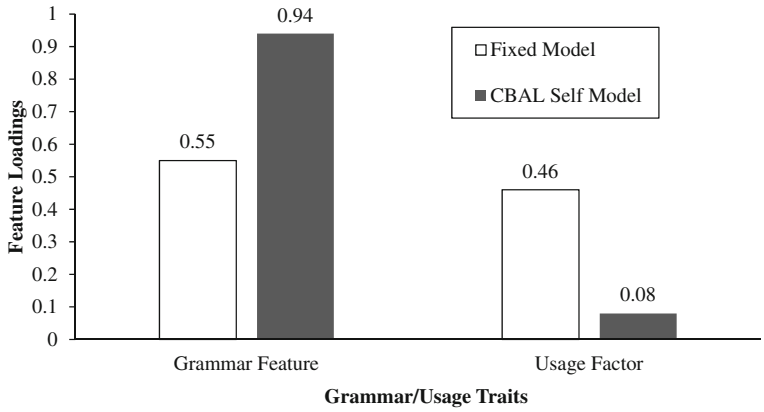

Figure 3 Relative loadings on the grammar and usage trait for the fixed model versus for a model trained on the CBAL essay corpus.

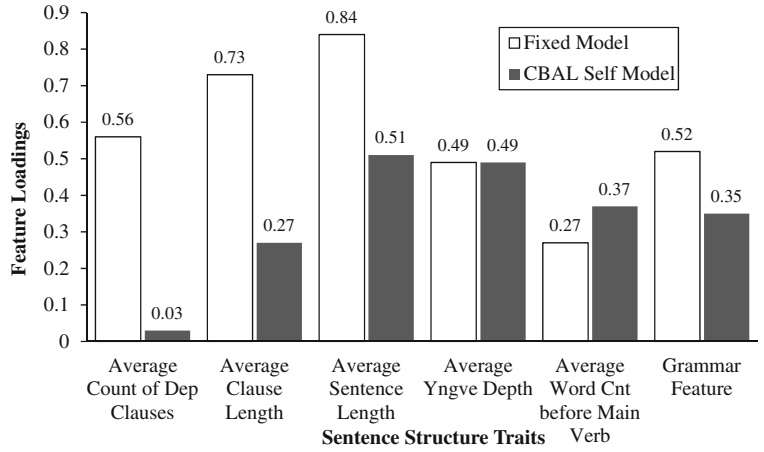

Figure 4 Relative loadings on the sentence structure trait for the fixed model versus for a model trained on the CBAL essay corpus.

Table 7 Correlations Between Trait Scores for the Fixed and Self Models on HiSET Essay Data

Trait	Correlation
Formality	1.00
Vocabulary length	1.00
Vocabulary frequency	1.00
Organization	.99
Sentence structure	.99
Sentence length	.99
Sentence complexity	1.00
Grammar and usage	1.00
Conventionality	1.00
Mechanics	1.00
Narrativity	.97
Contextualization	.97
Dialog	.97
Cohesion	1.00
Stance taking	1.00
Interactivity	1.00
Concreteness	.98

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 8 Correlations Between Trait Scores for the Fixed and Self Models on TOEFL Massive Open Online Course Essay Data

Trait	Correlation
Formality	.96
Vocabulary length	.98
Vocabulary frequency	1.00
Organization	1.00
Sentence structure	.89
Sentence length	.99
Sentence complexity	1.00
Grammar and usage	1.00
Conventionality	1.00
Mechanics	.93
Narrativity	.66
Contextualization	.67
Dialog	.96
Cohesion	1.00
Stance taking	1.00
Interactivity	.96
Concreteness	1.00

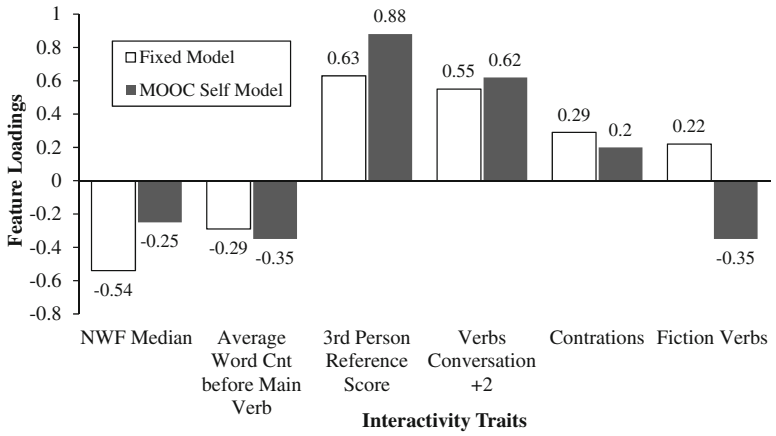

Figure 5 Relative loadings on the interactivity trait for the fixed model versus for a model trained on the TOEFL massive online open course corpus.

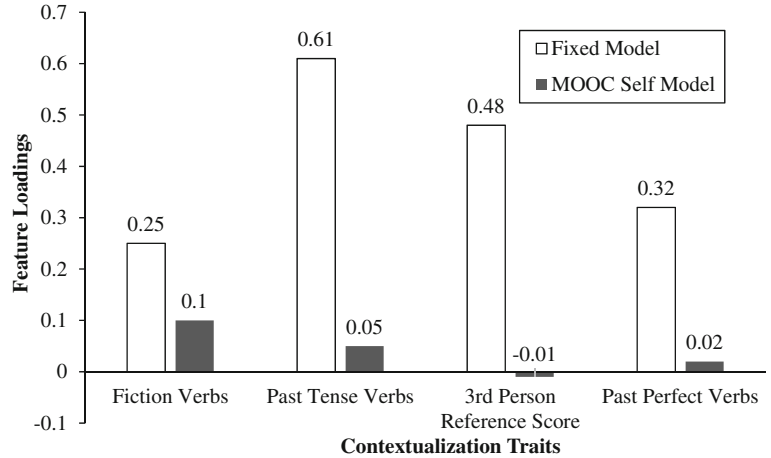

Figure 6 Relative loadings on the contextualization trait for the fixed model versus for a model trained on the TOEFL massive online open course corpus.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

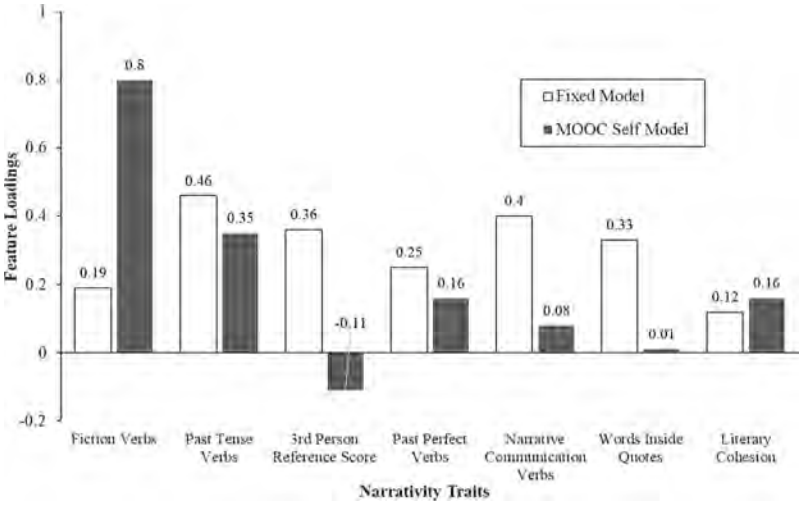

Figure 7 Relative loadings on the narrativity trait for the fixed model versus for a model trained on the TOEFL massive online open course corpus.

Table 9 Correlations Between Trait Scores for the Fixed and Self Models on TOEFL Essay Data

Trait	Correlation
Formality	.99
Vocabulary length	1.00
Vocabulary frequency	1.00
Organization	.99
Sentence structure	.98
Sentence length	.93
Sentence complexity	.99
Grammar and usage	1.00
Conventionality	1.00
Mechanics	.99
Narrativity	.98
Contextualization	1.00
Dialog	.97
Cohesion	1.00
Stance taking	1.00
Interactivity	1.00
Concreteness	.89

Model Performance on the TOEFL Essay Corpus

When we applied the fixed Criterion model to the TOEFL corpus, the fit was lower than what we observed for classroom Criterion essays (CFI = .861; RMSEA = .0730). However, when we retrained the model on the TOEFL data, the fit was acceptable, though slightly lower than it was on the original Criterion training set (CFI = .918; RMSEA = .0584). Except for the concreteness trait, $r = .89$, correlations between corresponding trait scores were consistently greater than .90 (see Table 9).

When we examined the factor loadings for concreteness, we found that the weight for the proportion of highly concrete words was much lower than it was for the fixed model, though the weight for the proportion of highly imageable words was similar (see Figure 8).

Model Performance on the GRE Essay Corpus

When we applied the fixed Criterion model to the GRE corpus, the fit was lower than what we observed for the classroom Criterion essays (CFI = .872; RMSEA = .0674). However, when we retrained the model on the GRE data, the fit

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

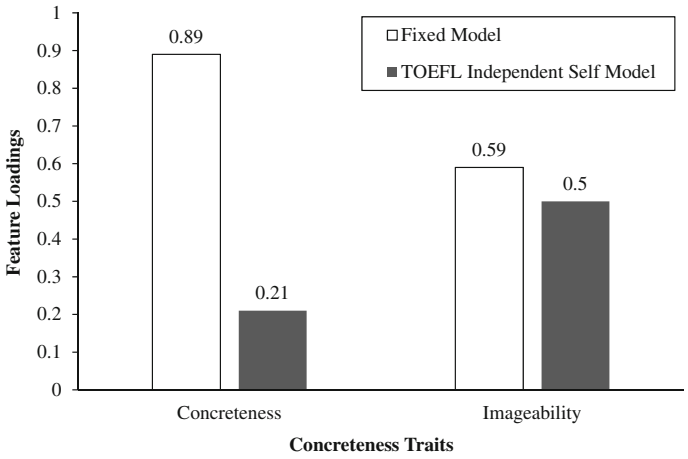

Figure 8 Relative loadings on the concreteness trait for the fixed model versus for a model trained on the TOEFL essay corpus.

Table 10 Correlations Between Trait Scores for the Fixed and Self Models on GRE Essay Data

Trait	Correlation
Formality	.97
Vocabulary length	1.00
Vocabulary frequency	1.00
Organization	1.00
Sentence structure	.99
Sentence length	.94
Sentence complexity	1.00
Grammar and usage	1.00
Conventionality	1.00
Mechanics	.97
Narrativity	.97
Contextualization	.98
Dialog	.97
Cohesion	1.00
Stance taking	.99
Interactivity	.99
Concreteness	.96

was acceptable and only slightly lower than it was on the original Criterion training set (CFI = .924; RMSEA = .0542). Correlations between corresponding trait scores were consistently greater than .94 (see Table 10).

Model Performance on the Institute for Educational Science-Funded Writing Achievement Study Lower Undergraduate Corpus

When we applied the fixed Criterion model to the GRE corpus, the fit was lower than what we observed for classroom Criterion essays (CFI = .723; RMSEA = .111). However, when we retrained the model on the GRE data, the fit was better, though still below acceptable levels (CFI = .868; RMSEA = .0761). Correlations between corresponding trait scores were all greater than .90, and most were greater than .98 (see Table 11).

Model Performance on the Michigan Corpus of Upper-Level Student Papers

When we applied the fixed Criterion model to the MICUSP, the fit was far lower than what we observed for classroom Criterion essays (CFI = .657; RMSEA = .127). When we retrained the model on the IES Writing Achievement Study data, the f**i**t improved, though it was still below an acceptable level of f**i**t and much lower than it was on the original Criterion training set (CFI = .779; RMSEA = .105). Three traits had correlations less than .90 between the fixed and self models:

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 11 Correlations Between Trait Scores for the Fixed and Self Models on Institute for Educational Science-Funded Writing Achievement Study College Essay Data

Trait	Correlation
Formality	.91
Vocabulary length	.99
Vocabulary frequency	.99
Organization	.98
Sentence structure	1.00
Sentence length	.91
Sentence complexity	1.00
Grammar and usage	.97
Conventionality	1.00
Mechanics	1.00
Narrativity	.92
Contextualization	1.00
Dialog	.91
Cohesion	1.00
Stance taking	1.00
Interactivity	.99
Concreteness	.99

Table 12 Correlations Between Trait Scores for the Fixed and Self Models on the Michigan Corpus of Upper-Level Student Papers Essay Data

Trait	Correlation
Formality	.94
Vocabulary length	.99
Vocabulary frequency	1.00
Organization	.99
Sentence structure	1.00
Sentence length	.99
Sentence complexity	1.00
Grammar and usage	.71
Conventionality	.98
Mechanics	.76
Narrativity	−.45
Contextualization	1.00
Dialog	.97
Cohesion	1.00
Stance taking	.98
Interactivity	.970
Concreteness	1.00

narrativity, $r = -.45$, grammar and usage, $r = .71$, and mechanics, $r = .76$. All other traits had correlations greater than .93 between the fixed and self models (see Table 12).

When we examined the relative weights for narrativity in the self model, the features for third person pronouns, for the two dialog features (narrative communication verbs and words inside quotes), and for fiction verbs were negative, which was the opposite sign than what was observed in the fixed model (see Figure 8). When we examined the relative weights for grammar and usage, the grammar feature had a near-zero weight in the self model and was much lower than in the fixed model, whereas the usage feature had similar weights in both models (see Figure 9). When we examined the relative weights for mechanics, the grammaticality feature had lower weights and the grammar feature much higher weights in the self model compared to the fixed model (see Figures 10 and 11).

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

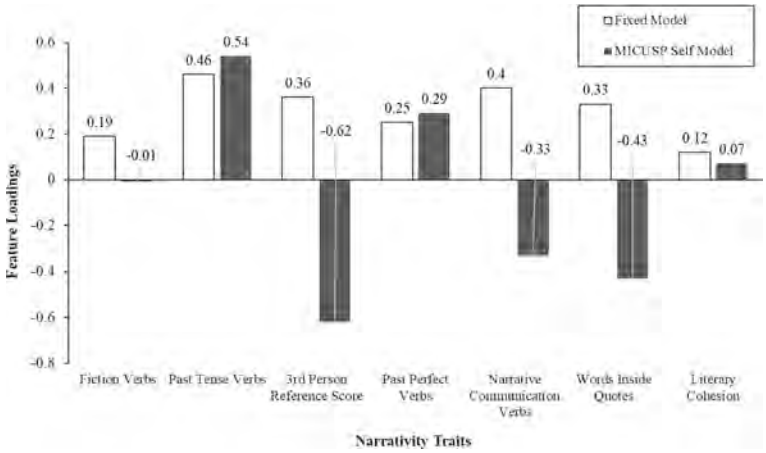

Figure 9 Relative loadings on the narrativity trait for the fixed model versus for a model trained on the TOEFL essay corpus.

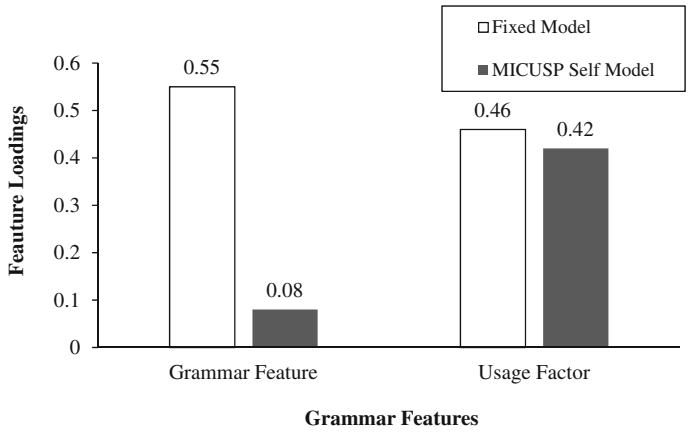

Figure 10 Relative loadings on the grammar and usage trait for the fixed model versus for a model trained on the TOEFL essay corpus.

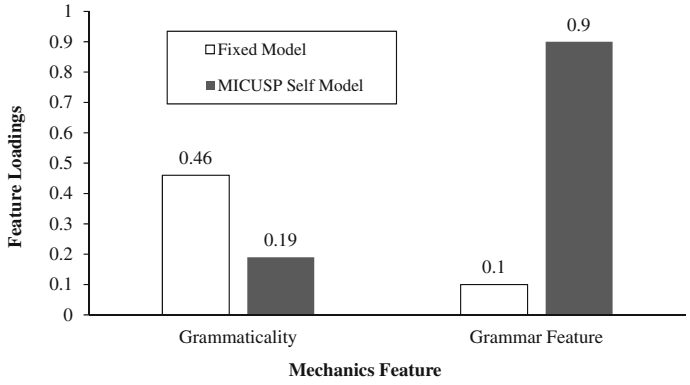

Figure 11 Relative loadings on the mechanics trait for the fixed model versus for a model trained on the TOEFL essay corpus.

Research Question 2: Discussion

Our results fell into two categories. The first category involves cases in which the fixed Criterion model had an acceptable fit at least on the self model and the traits on the fixed and self models were strongly correlated. This category included all the data sets that primarily consisted of essays: the original Criterion corpus, the Criterion human-scored corpus, the CBAL middle school corpus, the HiSET high school equivalency corpus, the TOEFL and TOEFL MOOC corpora, and

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 13 Distribution of Essay Length Across Samples (in Log Words)

Sample	*M* (in log words)	SD
Criterion classroom essays (training)	5.75	.58
Criterion classroom essays (test set)	5.75	.58
Criterion human scored	5.51	.62
CBAL	4.84	.87
HiSET	5.56	.48
TOEFL MOOC	5.54	.43
TOEFL	5.92	.42
GRE	5.80	.32
IES-funded Writing Achievement Study	6.53	.72
MICUSP	11.05	.91

Note. IES = Institute for Educational Sciences. MICUSP = Michigan Corpus of Upper-Level Student Papers. MOOC = massive online open course.

the GRE corpus. In each of these data sets, the self model had a better fit than the fixed model, though the difference in fit varied. For the operational Criterion corpus, there were only small differences in fit between the self and fixed models. The difference between the fixed and self models was somewhat larger for the human-scored Criterion corpus and much larger for the HiSET, CBAL middle school, GRE, TOEFL MOOC, and TOEFL data sets. The mismatch between the fixed and self models implies that the model structure was (at least mostly) appropriate, but either feature distributions or factor loadings differed. For all these data sets, given the high correlations between corresponding trait scores in the fixed and self models, it seems safe to conclude that the primary locus of difference is a difference in feature distributions, not a difference in factor loadings.

Two data sets in this category showed some mismatches between corresponding traits. In the CBAL data set, most of these differences were associated with sentence length features. In the TOEFL MOOC data set, specific features did not align with the interactivity, narrativity, and contextualization traits in a manner observed in other comparisons. These mismatches may have been caused by restricted sampling. The CBAL data were collected mostly from middle school students and represented primarily argumentative (and some informational, but no narrative) tasks. The TOEFL MOOC represented exactly two prompts (both argumentative in nature). Limited sampling, both of genres and of students, by definition increases the likelihood that a general model trained on a much broader spectrum of tasks or students will fail to fit in some fashion.

The second category involves cases in which neither the fixed nor the self model had an acceptable fit. Two data sets fell into this category: the IES Writing Achievement Study corpus of lower undergraduate papers and the MICUSP of upper undergraduate research papers. Both corpora are small (fewer than 1,000 papers) and consist of longer papers than the other data sets (see Table 13).

It appears that the trait model may break down when applied to texts of unrestricted length. It is worth noting, however, that for the IES Writing Achievement Study corpus and MICUSP, one of the larger differences between corresponding traits was the failure of several features assigned to the narrativity trait to load on that trait. Because upper undergraduate research papers are unlikely to be written as narratives, this failure may be entirely appropriate, especially considering the small size of the sample.

Overall, the 17-trait model appears to generalize well across a range of tasks and populations. For purposes in which scaling does not matter, such as predicting holistic essay scores, trait scores based on factor loadings derived from the original Criterion model may be applied reasonably to other data sets because they are typically correlated at .90 or better and often yielded correlations greater than .95.

Research Question 3: How Reliable Are These Traits?

Research Question 3: Results

As Tables 14 and 15 show, test–retest reliability was generally highest when we compared essays completed by the same student in the same month in response to Criterion topic library prompts from the same genre. Under these conditions, e-rater scores showed a reliability of .632 and essay length showed a reliability of .629. Four of the trait scores had higher

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 14 Test–Retest Reliabilities

Prompt/genre	Time span (months)	No. essay pairs	Test–retest reliability																			
			e-rater	Essay length	FRM	ORG	VLEN	VFRQ	SSTR	SLEN	SCPLX	GRM	CNV	MCH	NAR	CTX	DIA	COH	STN	INT	CNC	LT
Standard Criterion prompts																						
Same genre	1	16,970	.63	.63	.69	.67	.61	.45	.62	.63	.57	.41	.59	.57	.53	.50	.39	.32	.45	.51	.32	.23
	1[a]	13,782	.63	.62	.71	.63	.65	.50	.66	.67	.61	.45	.61	.60	.56	.53	.42	.35	.48	.57	.36	.22
	1[b]	1,302	.43	.42	.50	.42	.42	.23	.53	.54	.47	.28	.30	.50	.41	.41	.21	.14	.40	.27	.23	.32
	2	25,824	.61	.61	.68	.64	.59	.42	.60	.61	.55	.40	.57	.55	.52	.49	.38	.30	.43	.48	.29	.21
	2[a]	20,870	.61	.59	.69	.59	.63	.48	.64	.65	.60	.44	.59	.57	.55	.52	.41	.33	.46	.54	.33	.22
	2[b]	2,918	.41	.30	.47	.32	.41	.23	.50	.53	.42	.26	.51	.49	.39	.39	.20	.15	.36	.22	.24	.23
	12	52,428	.57	.57	.66	.59	.58	.39	.56	.57	.51	.37	.55	.52	.48	.44	.34	.26	.40	.46	.27	.17
Any genre	1	41,702	.61	.59	.62	.63	.52	.36	.57	.59	.52	.38	.57	.55	.18	.14	.17	.22	.15	.38	.17	.08
	2	55,448	.59	.57	.60	.60	.49	.32	.55	.57	.50	.37	.56	.54	.15	.13	.15	.19	.14	.33	.16	.15
	12	66,081	.54	.51	.56	.54	.46	.28	.49	.52	.44	.12	.52	.50	.09	.04	.11	.16	.10	.27	.33	.12
All prompts/ any genre	1	124,161	.52	.50	.58	.51	.48	.38	.51	.54	.47	.34	.53	.52	.19	.17	.17	.22	.18	.41	.15	.14
	2	155,215	.50	.48	.56	.49	.46	.35	.50	.52	.45	.33	.52	.50	.16	.14	.15	.20	.16	.38	.15	.13
	12	139,603	.47	.44	.53	.46	.43	.32	.44	.47	.40	.30	.48	.47	.12	.08	.13	.16	.13	.32	.13	.11

Note. CNC = concreteness. CNV = conventionality. COH = cohesion. CTX = contextualization. DIA = dialog. FRM = formality. GRM = grammar and usage. INT = interactivity. LT = lexical tightness. MCH = mechanics. NAR = narrativity. ORG = organization. SCPLX = sentence complexity. SLEN = sentence length. SSTR = sentence structure. VFRQ = vocabulary frequency. VLEN = vocabulary length. [a]Both >5.3 log words. [b]Both <5.3 log words.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 15 Correlations Between Trait Scores and e-rater Scores on One Criterion Essay With Trait Scores on a Second Essay Written in the Same Month in the Same Genre

Trait	Trait correlation	e-rater correlation
Formality	.69	.31
Vocabulary length	.61	.19
Vocabulary frequency	.45	−.13
Organization	.67	.58
Sentence structure	.62	.05
Sentence length	.63	−.04
Sentence complexity	.57	.14
Grammar and usage	.41	.21
Conventionality	.59	.16
Mechanics	.57	.11
Narrativity	.53	.13
Contextualization	.51	.11
Dialog	.39	.12
Cohesion	.32	−.05
Stance taking	.45	−.08
Interactivity	.51	−.07
Concreteness	.32	.09
Lexical tightness[a]	.23	.08

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

or approximately the same reliability as e-rater: organization, $r = .67$, formality, $r = .60$, vocabulary length, $r = .61$, and sentence length, $r = .66$. Six more traits had reliabilities greater than .50: conventionality, $r = .59$, mechanics, $r = .57$, sentence complexity, $r = .57$, narrativity, $r = .50$, contextualization, $r = .50$, and interactivity, $r = .51$. Vocabulary frequency had a reliability of .45, grammar and usage had a reliability of .41, stance taking had a reliability of .45, dialog had a reliability of .39, cohesion had a reliability of .32, and concreteness had a reliability of .32. The lexical tightness feature had a reliability of .23.

Increasing the time span between essays decreased reliability. For instance, e-rater reliability decreased from .63 to .61 when essays were sampled in a 2-month span, and reliability decreased to .57 when we examined pairs of essays from the beginning and the end of the school year. All the trait scores showed a similar pattern, with reliability decreasing at greater time spans between the essay pairs sampled. Allowing paired essays to differ in genre also decreased reliability. For instance, e-rater reliability decreased from .63 to .61 if genre differences were ignored. Most of the traits showed even larger decreases in reliability when genre was ignored. Sampling from the full range of prompts, including teacher-created prompts, also decreased reliability. For example, e-rater reliability for any pair sampled from the same student within the same month decreased from .61 to .52. The trait scores also showed similar patterns of decreasing reliability, though many of the decreases were small. For instance, conventionality had a reliability of .59 when essay pairs were sampled for the same student in the same genre in the same month from Criterion library prompts; however, reliability was .53 when essay pairs were sampled for the same student in the same month for all prompts, without regard to genre.

However, the reliability of certain traits (stance taking, narrativity, contextualization, concreteness, cohesion, and lexical tightness) decreased far more than for the other traits if essay pairs were sampled across genres. For instance, the reliability of stance taking decreased from .45 when students were sampled in the same month and genre to only .15 if genre was ignored.

Finally, the length of the essay also affected reliability. When we sampled essays written by the same student within the same month to Criterion topic library prompts, but limited the sample to essays greater than 5.3 essay length on a log words scale, reliability increased for all the traits, except organization (the trait most strongly correlated with essay length). For this sample, we observed reliabilities of .71 for formality, .65 for vocabulary length, .50 for vocabulary frequency, .63 for organization (a rare decrease), .66 for sentence structure, .67 for sentence length, .45 for grammar and usage, .61 for conventionality, .60 for mechanics, .56 for narrativity, .53 for contextualization, .42 for dialog, .48 for stance taking, .57 for interactivity, and .35 for concreteness. Conversely, if we limited the sample to essays less than or equal to an essay length of 5.3 on a log words scale, reliability decreased (.50 for formality, .42 for vocabulary length, .23 for vocabulary frequency,

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 16 Regression of e-rater Scores on the Second Essay From e-rater Scores and Trait Scores on the First Essay in a Pair of Final Draf tCriterion Essays Written by the Same Student in the Same Month in the Same Genre

Trait	*B*	*SE*	*t*	*p*
Constant	2.343	0.059	39.84	<0.0001
e-rater score	0.448	0.014	32.33	<0.001
Formality	−0.119	0.024	−4.94	<0.001
Vocabulary length	0.098	0.018	5.36	<0.001
Vocabulary frequency	−0.038	0.013	−2.91	0.004
Organization	0.244	0.020	12.31	<0.001
Sentence structure	−0.163	0.070	−2.35	0.019
Sentence length	0.013	0.048	0.26	0.795
Sentence complexity	0.110	0.027	4.12	<0.001
Grammar and usage	0.101	0.014	7.13	<0.001
Conventionality	0.128	0.032	3.99	<0.001
Mechanics	−0.097	0.033	−2.90	0.004
Narrativity	−0.392	0.416	−0.94	0.346
Contextualization	0.265	0.304	0.87	0.384
Dialog	0.174	0.195	0.90	0.370
Cohesion	0.004	0.038	0.12	0.908
Stance taking	−0.028	0.009	−3.17	0.002
Interactivity	0.026	0.013	2.03	0.043
Concreteness	0.053	0.008	6.33	<0.001
Lexical tightness[a]	−0.050	0.008	−6.60	<0.001

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

.42 for organization, .53 for sentence structure, .54 for sentence length, .28 for grammar and usage, .30 for conventionality, .50 for mechanics, .40 for narrativity, .40 for contextualization, .21 for dialog, .40 for stance taking, .27 for interactivity, and .23 for concreteness).

It should also be noted that the observed decreases appear to combine traits. For instance, allowing for a broader range of genres or accepting a longer span between the first and last essays in a pair corresponds to successively larger decreases in reliability.

Next, we examined the difference between trait-by-trait correlations and e-rater by trait correlations for students who responded to Criterion topic library prompts in the same genre in the same month of the school year. If the trait model is valid, we would expect trait scores on the first essay to be more strongly associated with trait scores on the second essay than they are with e-rater scores on the first essay. The results of this comparison are shown in Table 15. These results indicate that even for the organization trait, which has a high correlation with essay length, the trait score on the first essay is more reliably associated with the trait score on the second essay than are e-rater scores or essay length.

Next, we regressed e-rater scores on the second essay against e-rater and trait scores on the first essay. This regression indicated that a significant relationship existed, $F(10, 16{,}950) = 649.9$, $p < .001$, $R^2 = .42$. Several of the trait scores on the first essay were significant predictors of e-rater scores on the second essay, above and beyond e-rater scores on the first essay. The results of the regression are shown in Table 16.[4] The formality, vocabulary length, vocabulary frequency, organization, sentence structure, sentence complexity, grammar and usage, conventionality, mechanics, stance taking, interactivity, and concreteness trait scores on the first essay all accounted for a significant proportion of the variance in e-rater scores on the second essays above and beyond that accounted for by the e-rater score on the first essay.[5]

Finally, we analyzed the contribution that trait scores made above human score on one essay to predict the human score on another essay. These data were drawn from a CBAL data set in which students wrote two essays from one of two genres (policy argument or a cost–benefit analysis) within a 3-week period. Table 17 lists the genres and prompts; Table 18 shows the distribution of essays by genre and prompt; Table 19 shows the results of the regression. The vocabulary length, organization, sentence complexity, grammar and usage, and conventionality trait scores on the first essay in each pair all contributed significant independent variance to the prediction of the human score for the second essay in the pair, above and beyond the human score on the first essay.[6]

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 17 Genres and Topics in the CBAL Genre Study Data Set

Genre	Prompt
Policy argument	Should the United States ban advertisements to children under 12 years of age?
	Should schools pay students cash for getting good grades?
	Should parents set limits on students' use of social networking sites?
Cost–benefit analysis	What should our class choose as a service-learning project?
	How should our school spend a generous gift?
	What theme should our school choose for a culture fair?

Table 18 Distribution of Topics Across Essay Pairs in the CABL Genre Study Data Set

First prompt	Second prompt	*n*
Ban Ads	Cash for Grades	332
Ban Ads	Social Networking	264
Cash for Grades	Social Networking	545
Generous Gift	Service Learning	313
Culture Fair	Service Learning	309
Culture Fair	Generous Gift	627

Table 19 Regressions of Human Scores on One Essay From Human Scores and Trait Scores on the Other Essay in Pairs of CBAL Essays Written by the Same Student in the Same Month in the Same Genre

Trait	*B*	*SE*	*t*	*p*
Constant	3.835	0.118	32.62	<0.001
Human scored	0.253	0.022	11.51	<0.001
Formality	−0.101	0.083	−1.21	0.225
Vocabulary length	0.283	0.059	4.78	<0.001
Vocabulary frequency	−0.033	0.039	−0.85	0.397
Organization	0.678	0.062	10.99	<0.001
Sentence structure	−0.372	0.203	−1.83	0.067
Sentence length	0.224	0.150	1.49	0.136
Sentence complexity	0.165	0.080	2.07	0.039
Grammar and usage	0.157	0.042	3.78	<0.001
Conventionality	0.249	0.089	2.79	0.005
Mechanics	−0.058	0.091	−0.64	0.525
Narrativity	−0.463	1.708	−0.27	0.787
Contextualization	0.28	1.047	0.27	0.791
Dialog	0.304	1.113	0.27	0.785
Cohesion	0.094	0.183	0.51	0.607
Stance taking	0.062	0.030	0.88	0.380
Interactivity	0.001	0.038	0.02	0.982
Concreteness	−0.015	0.029	0.53	0.596

Research Question 3: Discussion

Our results indicate that some of the trait scores were more reliable than e-rater scores and that most of them had test–retest reliabilities greater than .5 in the 1-month condition in Table 14. Note that the test–retest reliability estimates we report should be viewed as lower bounds, because they are likely to reflect reductions in reliability due to differences among prompts within the same genre. In every case, the trait score was more strongly correlated with the retest performance on the same trait than with e-rater scores, indicating that the traits captured distinct variance from the unidimensional model captured by e-rater. Furthermore, when we directly examined the contribution of trait scores to predicting performance on the second essay of a pair, a majority of the traits were significant predictors above and beyond the e-rater or human scores.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 20 Submissions by Genre for Criterion Topic Library Prompts

Genre	No. of submissions
Cause and effect	26,268
Compare/contrast	39,075
Descriptive	99,057
Expository	82,082
Persuasive	187,103
Narrative	129,075
Process	30,817

Table 21 Trait Means and Standard Deviations by Genre

Trait	Persuasive	Cause/effect	Expository	Compare/contrast	Descriptive	Process	Narrative
Formality	0.05 (1.00)	−0.32 (0.85)	0.09 (0.88)	−0.43 (0.82)	−0.41 (0.87)	−0.74 (0.80)	−0.71 (0.79)
Vocabulary length	0.04 (1.00)	−0.20 (0.92)	0.05 (0.91)	−0.22 (0.80)	−0.32 (0.95)	−0.81 (0.87)	−0.78 (0.76)
Vocabulary frequency	0.18 (0.90)	0.46 (0.80)	0.25 (0.90)	0.36 (0.86)	0.12 (0.90)	0.30 (1.03)	0.62 (0.87)
Organization	−0.03 (0.90)	−0.44 (0.84)	0.14 (0.83)	−0.25 (0.85)	−0.23 (0.90)	−0.45 (0.87)	0.09 (1.04)
Sentence structure	0.19 (0.97)	−0.01 (1.01)	0.20 (0.89)	−0.04 (1.04)	−0.16 (0.98)	−0.20 (1.05)	−0.20 (1.12)
Sentence length	0.22 (0.98)	0.04 (1.04)	0.16 (0.92)	−0.06 (1.08)	−0.20 (0.98)	−0.16 (1.06)	−0.09 (1.16)
Sentence complexity	0.05 (0.96)	0.09 (0.97)	0.14 (0.90)	−0.03 (1.05)	−0.09 (1.00)	−0.20 (0.99)	−0.27 (1.03)
Grammar and usage	0.09 (1.01)	0.07 (1.00)	0.19 (0.92)	0.06 (1.03)	−0.00 (1.00)	−0.09 (1.05)	−0.10 (0.97)
Conventionality	0.21 (0.93)	0.15 (1.05)	0.26 (0.89)	−0.06 (1.08)	0.09 (0.98)	0.02 (1.05)	0.12 (0.98)
Mechanics	0.20 (0.91)	0.22 (1.04)	0.27 (0.88)	−0.02 (1.05)	0.12 (0.97)	0.07 (1.02)	0.23 (0.98)
Narrativity	−0.64 (0.72)	−0.14 (0.92)	−0.34 (0.79)	−0.46 (0.76)	−0.22 (0.87)	−0.45 (0.89)	0.98 (0.91)
Contextualization	−0.67 (0.71)	−0.15 (0.85)	−0.29 (0.84)	−0.52 (0.83)	−0.10 (0.90)	−0.51 (0.81)	0.96 (0.80)
Dialog	−0.36 (0.81)	−0.05 (1.05)	−0.27 (0.80)	−0.22 (0.84)	−0.38 (0.88)	−0.21 (0.94)	0.68 (1.11)
Cohesion	0.17 (0.92)	−0.05 (0.94)	0.00 (0.93)	0.28 (1.02)	−0.20 (1.01)	0.26 (0.95)	−0.46 (1.00)
Stance taking	0.70 (0.95)	0.63 (1.23)	0.26 (0.90)	−0.13 (0.93)	−0.29 (0.81)	−0.05 (0.85)	−0.64 (0.70)
Interactivity	0.02 (0.86)	0.30 (0.82)	0.22 (0.86)	0.26 (0.85)	0.27 (0.84)	0.39 (0.74)	0.88 (0.70)
Concreteness	−0.29 (0.96)	0.09 (1.12)	−0.43 (0.94)	−0.28 (1.00)	0.43 (0.93)	0.32 (1.04)	0.54 (0.80)

Note. Standard deviations are in parentheses.

In the Criterion e-rater-based analysis, 12 traits were significant predictors. In the CBAL human score–based analysis, with a much smaller sample size, five traits were significant predictors. Overall, similar patterns were obtained. The trait model improves score prediction from one occasion to the next, though in either analysis, certain traits, especially narrativity and its subordinate traits, do not have significant connections with the score.

Overall, these results support the use of trait scores to profile student performance, because the pattern of trait scores contains information absent from overall, holistic scores. However, the results also indicate that genre differences and task type affect reliability. If we ignore genre differences and include a variety of teacher-created prompts in the mix, we might observe considerable reductions in the reliability of both e-rater and many trait scores. Thus, in future work, it may be critical to model genre variation in teacher-created assignments, because our data suggest that genre may account for considerable within-person variation.

Research Question 4: How Are the Traits Related to Differences in Genre?

Research Question 4: Results

General Results

The Criterion topic library contains seven genres, or modes of writing: cause and effect, compare/contrast, descriptive, expository, persuasive, narrative, and process. Table 20 shows the number of submissions per genre in the K–12 Criterion corpus.

Genre one-way ANOVAs were run for genre differences in trait scores in the Criterion topic library data. All 12 traits plus the informational cohesion and lexical tightness features returned results that indicate significant differences in genre means (see Table 21 for means and standard deviations). Tukey–Kramer post hoc tests revealed that almost all genres were

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 22 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Formality Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.11**	−0.14	−0.09	−0.14
	Descriptive	−0.09**	−0.12	−0.08	−0.12
	Expository	0.4**	0.38	0.42	0.47
	Narrative	−0.39**	−0.42	−0.38	−0.48
	Persuasive	0.36**	0.34	0.38	0.38
	Process	−0.42**	−0.45	−0.40	−0.52
Compare/contrast	Descriptive	0.01	0.00	0.03	0.02
	Expository	0.51**	0.50	0.53	0.61
	Narrative	−0.28**	−0.30	−0.27	−0.35
	Persuasive	0.47**	0.46	0.49	0.52
	Process	−0.31**	−0.33	−0.29	−0.39
Descriptive	Expository	0.5**	0.49	0.51	0.57
	Narrative	−0.29**	−0.31	−0.29	−0.36
	Persuasive	0.45**	0.45	0.47	0.49
	Process	−0.32**	−0.35	−0.31	−0.39
Expository	Narrative	−0.8**	−0.81	−0.79	−0.96
	Persuasive	−0.04**	−0.05	−0.03	−0.05
	Process	−0.82**	−0.85	−0.81	−0.99
Narrative	Persuasive	0.75**	0.75	0.77	−0.84
	Process	−0.02**	−0.05	−0.01	−0.04
Persuasive	Process	−0.78**	−0.80	−0.77	−0.87

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

significantly different at $p < .001$. Examination of effect sizes (Cohen's *d*) indicated that mean differences in traits by genre ranged from very small to extremely large (see Tables 22–39).[7]

Detailed Results by Trait

Formality. A one-way ANOVA indicated significant genre effects, $F(6, 569{,}377) = 12{,}129.58$, $p < .001$. Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with one exception: the contrast between compare/contrast and descriptive, $p = .071$. Expository essays had the highest mean formality scores, whereas process essays had the lowest formality scores (see Table 22). As seen in Table 22, the smallest effect sizes were for the contrasts narrative/process and expository/persuasive. The largest effect sizes were observed for the contrasts expository/process, expository/narrative persuasive/process, and narrative/persuasive. Overall, persuasive and expository essays were high on formality; process essays and narrative essays were low.

Vocabulary length. A one-way ANOVA indicated significant genre effects, $F(6, 569{,}377) = 13{,}171.83$, $p < .001$. Post hoc comparisons with a Tukey–Kramer correction showed that most contrasts were statistically significant, $ps < .05$, with the exceptions of cause and effect/compare/contrast, $p = .067$, and expository/persuasive, $p = .67$. Expository and persuasive essays demonstrated the highest mean vocabulary length scores, whereas process essays demonstrated the lowest (see Table 21). The smallest effect size was observed for the contrast between narrative and process, whereas the largest effect size was observed for the contrast between expository and narrative (see Table 23). Overall, expository and persuasive essays were associated with longer vocabulary length scores, and narrative and process essays were shorter on the vocabulary length trait.

Vocabulary frequency. Descriptive essays had the lowest mean vocabulary frequency scores, whereas narratives had the highest, $F(6, 569{,}377) = 3{,}917.88$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, $ps < .05$. Table 24 shows the effect sizes (Cohen's *d*). The smallest effect size was observed for the contrast expository/persuasive, and the largest effect size was observed for the contrast narrative/descriptive, $d = .594$. Overall, descriptive and persuasive essays were associated with low vocabulary frequency, and narratives were associated with high vocabulary frequency.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 23 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Vocabulary Length Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.02	−0.04	0.00	−0.02
	Descriptive	−0.12**	−0.14	−0.10	−0.14
	Expository	0.25**	0.22	0.27	0.27
	Narrative	−0.58**	−0.60	−0.56	−0.69
	Persuasive	0.24**	0.22	0.26	0.26
	Process	−0.61**	−0.64	−0.59	−0.69
Compare/contrast	Descriptive	−0.10**	−0.12	−0.08	−0.11
	Expository	0.27**	0.25	0.28	0.31
	Narrative	−0.56**	−0.58	−0.54	−0.72
	Persuasive	0.26**	0.25	0.28	0.29
	Process	−0.59**	−0.61	−0.57	−0.71
Descriptive	Expository	0.37**	0.35	0.38	0.42
	Narrative	−0.46**	−0.47	−0.45	−0.57
	Persuasive	0.36**	0.35	0.37	0.39
	Process	−0.49**	−0.51	−0.47	−0.57
Expository	Narrative	−0.83**	−0.84	−0.82	−0.99
	Persuasive	−0.01	−0.02	0.00	−0.01
	Process	−0.86**	−0.88	−0.84	−0.96
Narrative	Persuasive	0.82**	0.81	0.83	0.93
	Process	−0.03**	−0.05	−0.01	−0.04
Persuasive	Process	−0.85**	−0.87	−0.84	−0.92

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 24 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Vocabulary Frequency Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.10**	−0.12	−0.08	−0.12
	Descriptive	−0.34**	−0.35	−0.32	−0.40
	Expository	−0.21**	−0.23	−0.19	−0.25
	Narrative	0.16**	−0.18	−0.14	0.20
	Persuasive	−0.28**	−0.30	−0.26	−0.31
	Process	−0.16**	−0.18	−0.14	−0.17
Compare/contrast	Descriptive	−0.24**	−0.25	−0.22	−0.28
	Expository	−0.11**	−0.13	−0.09	−0.12
	Narrative	0.26**	0.24	0.27	0.32
	Persuasive	−0.18**	−0.19	−0.16	−0.20
	Process	−0.06**	−0.08	−0.04	−0.06
Descriptive	Expository	0.13**	0.12	0.14	0.14
	Narrative	0.50**	0.48	0.51	0.59
	Persuasive	0.06**	0.05	0.07	0.07
	Process	0.18**	0.16	0.19	0.18
Expository	Narrative	0.37**	0.36	0.38	0.44
	Persuasive	−0.07**	−0.08	−0.06	−0.08
	Process	0.05**	0.03	0.07	0.05
Narrative	Persuasive	−0.44**	−0.45	−0.43	−0.52
	Process	−0.32**	−0.34	−0.30	−0.35
Persuasive	Process	0.12**	0.10	0.13	0.12

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 25 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Organization Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.18**	−0.20	−0.16	−0.22
	Descriptive	0.03*	0.01	0.05	0.03
	Expository	0.39**	0.37	0.41	0.47
	Narrative	0.34**	0.32	0.36	0.36
	Persuasive	0.23**	0.21	0.25	0.25
	Process	−0.19**	−0.21	−0.17	−0.22
Compare/contrast	Descriptive	0.21**	0.19	0.23	0.24
	Expository	0.58**	0.56	0.60	0.69
	Narrative	0.52**	0.51	0.54	0.55
	Persuasive	0.41**	0.40	0.43	0.47
	Process	−0.01	−0.03	0.02	−0.01
Descriptive	Expository	0.37**	0.35	0.38	0.42
	Narrative	0.31**	0.30	0.33	0.32
	Persuasive	0.20**	0.19	0.21	0.23
	Process	−0.22**	−0.23	−0.20	−0.24
Expository	Narrative	−0.05**	−0.07	−0.04	−0.06
	Persuasive	−0.16**	−0.18	−0.15	−0.19
	Process	−0.58**	−0.60	−0.56	−0.69
Narrative	Persuasive	−0.11**	−0.12	−0.10	−0.11
	Process	−0.53**	−0.55	−0.51	−0.55
Persuasive	Process	−0.42**	−0.44	−0.40	−0.47

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 26 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Sentence Structure Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.04**	−0.06	−0.01	−0.04
	Descriptive	−0.16**	−0.18	−0.14	−0.22
	Expository	0.21**	0.19	0.23	0.18
	Narrative	−0.19**	−0.21	−0.17	−0.18
	Persuasive	0.20**	0.18	0.22	0.20
	Process	−0.20**	−0.22	−0.17	−0.19
Compare/contrast	Descriptive	−0.12**	−0.14	−0.10	−0.12
	Expository	0.25**	0.23	0.26	0.25
	Narrative	−0.16**	−0.17	−0.14	−0.14
	Persuasive	0.24**	0.22	0.25	0.23
	Process	−0.16**	−0.18	−0.13	−0.15
Descriptive	Expository	0.37**	0.38	0.38	0.39
	Narrative	−0.04**	−0.05	−0.02	−0.03
	Persuasive	0.36**	0.34	0.37	0.36
	Process	−0.04**	−0.06	−0.02	−0.04
Expository	Narrative	−0.40**	−0.42	−0.39	−0.40
	Persuasive	−0.01	−0.02	0.00	−0.01
	Process	−0.40**	−0.43	−0.38	−0.42
Narrative	Persuasive	0.39**	0.38	0.40	0.37
	Process	0.00	−0.02	0.02	0.00
Persuasive	Process	−0.39**	−0.41	−0.37	−0.39

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 27 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Sentence Length Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.10**	−0.12	−0.07	−0.09
	Descriptive	−0.24**	−0.26	−0.21	−0.23
	Expository	0.12**	0.09	0.14	0.12
	Narrative	−0.14**	−0.16	−0.11	−0.12
	Persuasive	0.18**	0.16	0.20	0.18
	Process	−0.20**	−0.22	−0.17	−0.19
Compare/contrast	Descriptive	−0.14**	−0.16	−0.12	−0.14
	Expository	0.22**	0.20	0.24	0.22
	Narrative	−0.04**	−0.06	−0.02	−0.03
	Persuasive	0.28**	0.26	0.29	0.27
	Process	−0.10**	−0.12	−0.07	−0.09
Descriptive	Expository	0.36**	0.34	0.37	0.38
	Narrative	0.10**	0.09	0.11	0.09
	Persuasive	0.41**	0.40	0.43	0.42
	Process	0.04**	0.02	0.06	0.04
Expository	Narrative	−0.26**	−0.27	−0.24	−0.25
	Persuasive	0.06**	0.04	0.07	0.06
	Process	−0.32**	−0.34	−0.30	0.32
Narrative	Persuasive	0.31**	0.30	0.32	0.29
	Process	−0.06**	−0.08	−0.04	−0.05
Persuasive	Process	−0.37**	−0.39	−0.35	−0.36

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 28 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Sentence Complexity Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	0.06**	0.04	0.09	−0.06
	Descriptive	0.00	−0.02	0.02	0.00
	Expository	0.23**	0.20	0.25	0.24
	Narrative	−0.18**	−0.20	−0.16	−0.18
	Persuasive	0.13**	0.11	0.15	0.14
	Process	−0.11**	−0.13	−0.18	−0.11
Compare/contrast	Descriptive	−0.06**	−0.08	−0.04	−0.06
	Expository	0.17**	0.15	0.19	0.17
	Narrative	−0.24**	−0.26	−0.23	−0.23
	Persuasive	0.07**	0.05	0.09	0.07
	Process	−0.17**	−0.19	−0.15	−0.17
Descriptive	Expository	0.23**	0.21	0.24	0.24
	Narrative	−0.18**	−0.19	−0.17	−0.18
	Persuasive	0.13**	0.12	0.15	0.14
	Process	−0.11**	−0.13	−0.09	−0.11
Expository	Narrative	−0.41**	−0.42	−0.40	−0.42
	Persuasive	−0.09**	−0.11	−0.08	−0.10
	Process	−0.34**	−0.36	−0.32	−0.36
Narrative	Persuasive	0.31**	0.30	0.33	0.32
	Process	0.07**	0.05	0.09	0.07
Persuasive	Process	−0.24**	−0.26	−0.22	−0.25

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 29 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Grammar and Usage Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.01	−0.03	0.01	−0.01
	Descriptive	−0.07**	−0.09	−0.05	−0.07
	Expository	0.12**	0.10	0.14	0.12
	Narrative	−0.16**	−0.18	−0.14	−0.17
	Persuasive	0.03*	0.00	0.05	0.03
	Process	−0.16**	−0.18	−0.14	−0.16
Compare/contrast	Descriptive	−0.06**	−0.08	−0.04	−0.06
	Expository	0.13**	0.11	0.15	0.13
	Narrative	−0.15**	−0.17	−0.14	−0.15
	Persuasive	0.04**	0.02	0.05	0.03
	Process	−0.15**	−0.17	−0.13	−0.14
Descriptive	Expository	0.19**	0.17	0.20	0.20
	Narrative	−0.09**	−0.11	−0.08	−0.09
	Persuasive	0.10**	0.08	0.11	0.10
	Process	−0.09**	−0.11	−0.07	−0.09
Expository	Narrative	−0.28**	−0.29	−0.27	−0.30
	Persuasive	−0.09**	−0.11	−0.08	−0.10
	Process	−0.28**	−0.30	−0.26	−0.28
Narrative	Persuasive	0.19**	0.18	0.20	0.19
	Process	0.00	−0.02	0.02	0.00
Persuasive	Process	−0.19**	−0.20	−0.17	−0.18

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 30 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Conventionality Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.22**	−0.24	−0.19	−0.20
	Descriptive	−0.06**	−0.08	−0.04	−0.06
	Expository	0.10**	0.08	0.12	0.10
	Narrative	−0.03**	−0.05	−0.01	−0.03
	Persuasive	0.05**	0.03	0.07	0.06
	Process	−0.13**	−0.16	−0.11	−0.13
Compare/contrast	Descriptive	0.16**	0.14	0.17	0.15
	Expository	0.32**	0.30	0.34	0.32
	Narrative	0.18**	0.17	0.20	0.18
	Persuasive	0.27**	0.25	0.29	0.27
	Process	0.08**	0.06	0.11	−0.08
Descriptive	Expository	0.16**	0.15	0.18	0.17
	Narrative	0.03**	0.01	0.04	0.03
	Persuasive	0.11**	0.10	0.13	0.12
	Process	−0.07**	−0.09	−0.05	−0.07
Expository	Narrative	−0.13**	−0.15	−0.12	−0.14
	Persuasive	−0.05**	−0.06	−0.03	−0.05
	Process	−0.23**	−0.25	−0.21	−0.24
Narrative	Persuasive	0.09**	0.08	0.10	0.09
	Process	−0.10**	−0.12	−0.08	−0.10
Persuasive	Process	−0.19**	−0.21	−0.17	−0.19

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 31 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Mechanics Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.23**	−0.26	−0.21	−0.22
	Descriptive	−0.10**	−0.12	−0.08	−0.10
	Expository	0.05**	0.03	0.01	−0.06
	Narrative	0.01	−0.01	0.03	0.01
	Persuasive	−0.02*	−0.04	0.00	−0.02
	Process	−0.15**	−0.17	−0.12	−0.14
Compare/contrast	Descriptive	0.13**	0.12	0.15	0.13
	Expository	0.29**	0.27	0.31	0.30
	Narrative	0.24**	0.23	0.26	0.24
	Persuasive	0.21**	0.20	0.23	0.22
	Process	0.08**	0.06	0.11	−0.08
Descriptive	Expository	0.15**	0.14	0.17	0.16
	Narrative	0.11**	0.10	0.12	0.11
	Persuasive	0.08**	0.07	0.09	0.08
	Process	−0.05**	−0.07	−0.03	−0.05
Expository	Narrative	−0.04**	−0.06	−0.03	−0.05
	Persuasive	−0.08**	−0.09	−0.06	−0.08
	Process	−0.20**	−0.22	−0.18	−0.21
Narrative	Persuasive	−0.03**	−0.04	−0.02	−0.03
	Process	−0.16**	−0.18	−0.14	−0.16
Persuasive	Process	−0.13**	−0.15	−0.11	−0.13

Note. Post-hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 32 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Narrativity Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.32**	−0.34	−0.30	−0.38
	Descriptive	−0.08**	−0.10	−0.06	−0.09
	Expository	−0.20**	−0.22	−0.18	−0.23
	Narrative	1.13**	1.11	1.14	1.23
	Persuasive	−0.50**	−0.52	−0.48	−0.63
	Process	−0.31**	−0.33	−0.29	−0.34
Compare/contrast	Descriptive	0.24**	0.23	0.26	0.27
	Expository	0.12**	0.11	0.14	0.16
	Narrative	1.45**	1.43	1.46	1.72
	Persuasive	−0.18**	−0.19	−0.17	−0.24
	Process	0.01	−0.01	0.03	0.02
Descriptive	Expository	−0.12**	−0.13	−0.11	−0.14
	Narrative	1.20**	1.19	1.22	1.35
	Persuasive	−0.42**	−0.43	−0.41	−0.53
	Process	−0.23**	−0.25	−0.21	−0.26
Expository	Narrative	1.32**	1.31	1.34	1.55
	Persuasive	−0.30**	−0.31	−0.29	−0.40
	Process	−0.11**	−0.13	−0.09	−0.13
Narrative	Persuasive	−1.63**	−1.63	−1.62	−1.98
	Process	−1.43**	−1.45	−1.42	−1.59
Persuasive	Process	0.19**	0.18	0.21	0.24

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 33 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Contextualization Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.37**	−0.39	−0.35	−0.45
	Descriptive	0.05**	0.03	0.07	0.06
	Expository	−0.15**	−0.16	−0.13	−0.17
	Narrative	1.11**	1.09	1.12	1.34
	Persuasive	−0.52**	−0.54	−0.50	−0.64
	Process	−0.36**	−0.38	−0.34	−0.43
Compare/contrast	Descriptive	0.42**	0.41	0.44	0.48
	Expository	−0.15**	−0.16	−0.13	−0.27
	Narrative	1.48**	1.47	1.49	1.82
	Persuasive	−0.15**	−0.16	−0.13	−0.19
	Process	0.01	0.00	0.03	0.02
Descriptive	Expository	−0.20**	−0.21	−0.18	−0.23
	Narrative	1.06**	1.05	1.07	1.24
	Persuasive	−0.57**	−0.58	−0.56	−0.70
	Process	−0.41**	−0.42	−0.39	−0.48
Expository	Narrative	1.25**	1.24	1.27	1.53
	Persuasive	−0.37**	−0.38	−0.36	−0.48
	Process	−0.21**	−0.23	−0.19	−0.26
Narrative	Persuasive	−1.63**	−1.64	−1.62	−2.15
	Process	−1.47**	−1.48	−1.45	−1.81
Persuasive	Process	0.16**	0.15	0.18	0.21

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 34 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Dialog Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.16**	−0.18	−0.14	−0.17
	Descriptive	−0.22**	−0.24	−0.21	−0.23
	Expository	−0.21**	−0.23	−0.18	−0.23
	Narrative	0.74**	0.71	0.76	0.68
	Persuasive	−0.30**	−0.32	−0.28	−0.36
	Process	−0.16**	−0.18	−0.13	−0.16
Compare/contrast	Descriptive	−0.06**	−0.08	−0.05	−0.07
	Expository	−0.05**	−0.07	−0.03	−0.06
	Narrative	0.90**	0.88	0.91	0.91
	Persuasive	−0.14**	−0.16	−0.13	−0.17
	Process	0.01	−0.02	0.03	0.01
Descriptive	Expository	0.01	0.00	0.03	0.02
	Narrative	0.96**	0.95	0.97	0.96
	Persuasive	−0.08**	−0.09	−0.07	−0.09
	Process	0.07**	0.05	0.09	0.08
Expository	Narrative	0.95**	0.94	0.96	0.98
	Persuasive	−0.09**	−0.10	−0.08	−0.11
	Process	0.06**	0.04	0.08	0.07
Narrative	Persuasive	−1.04**	−1.05	−1.03	−1.07
	Process	−0.89**	−0.91	−0.87	−0.87
Persuasive	Process	0.15**	0.13	0.17	0.17

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 35 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Cohesion Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	0.33**	0.31	0.36	0.34
	Descriptive	−0.15**	−0.17	−0.13	−0.15
	Expository	0.06**	0.04	0.08	0.06
	Narrative	−0.41**	−0.43	−0.39	−0.42
	Persuasive	0.22**	0.20	0.24	0.23
	Process	0.32**	0.29	0.34	0.33
Compare/contrast	Descriptive	−0.48**	−0.50	−0.46	−0.49
	Expository	−0.28**	−0.30	−0.26	−0.28
	Narrative	−0.74**	−0.76	−0.73	−0.74
	Persuasive	−0.11**	−0.13	−0.09	−.011
	Process	−0.02	−0.04	0.01	−0.02
Descriptive	Expository	0.20**	0.19	0.22	0.21
	Narrative	−0.26**	−0.27	−0.25	−0.26
	Persuasive	0.37**	0.36	0.38	0.38
	Process	0.46**	0.44	0.48	0.47
Expository	Narrative	−0.47**	−0.48	−0.45	−0.48
	Persuasive	0.17**	0.15	0.18	0.18
	Process	0.26**	0.24	0.28	0.28
Narrative	Persuasive	0.63**	0.62	0.64	0.66
	Process	0.73**	0.71	0.74	0.74
Persuasive	Process	0.09**	0.07	0.11	0.10

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 36 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Stance Taking Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.77**	−0.79	−0.74	−0.70
	Descriptive	−0.92**	−0.94	−0.90	−0.88
	Expository	−0.37**	−0.39	−0.35	−0.34
	Narrative	−1.27**	−1.29	−1.25	−1.27
	Persuasive	0.07**	0.05	0.09	0.08
	Process	−0.68**	−0.71	−0.66	−0.65
Compare/contrast	Descriptive	−0.15**	−0.17	−0.14	−0.15
	Expository	0.40**	0.38	0.42	0.44
	Narrative	−0.50**	0.52	0.49	−0.61
	Persuasive	0.83**	0.82	0.85	0.89
	Process	0.08**	0.06	0.10	0.09
Descriptive	Expository	0.55**	0.54	0.57	0.65
	Narrative	−0.35**	−0.36	−0.34	−0.46
	Persuasive	0.99**	0.98	1.00	1.12
	Process	0.24**	0.22	0.25	0.28
Expository	Narrative	−0.90**	−0.91	−0.89	1.12
	Persuasive	0.44**	0.42	0.45	0.47
	Process	−0.32**	−0.34	−0.30	−0.36
Narrative	Persuasive	1.34**	1.33	1.35	1.60
	Process	0.58**	0.57	0.60	0.75
Persuasive	Process	−0.75**	−0.77	−0.74	−0.84

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 37 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Interactivity Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.04**	−0.06	−0.02	−0.05
	Descriptive	−0.03**	−0.04	−0.01	−0.03
	Expository	−0.08**	−0.10	−0.07	−0.10
	Narrative	0.57**	0.56	0.59	0.75
	Persuasive	−0.28**	−0.29	−0.26	−0.33
	Process	0.09**	0.06	0.11	0.11
Compare/contrast	Descriptive	0.02*	0.00	0.03	0.02
	Expository	−0.04**	−0.06	−0.02	−0.05
	Narrative	0.61**	0.60	0.63	0.79
	Persuasive	−0.24**	−0.25	−0.22	−0.28
	Process	0.13**	0.11	0.15	0.16
Descriptive	Expository	−0.06**	−0.07	−0.05	−0.07
	Narrative	0.60**	0.59	0.61	0.78
	Persuasive	−0.25**	−0.26	−0.24	−0.30
	Process	0.11**	0.09	0.13	0.14
Expository	Narrative	0.66**	0.65	0.67	0.84
	Persuasive	−0.19**	−0.20	−0.18	−0.23
	Process	0.17**	0.15	0.19	−0.21
Narrative	Persuasive	−0.85**	−0.86	−0.84	−1.09
	Process	−0.49**	−0.51	−0.47	−0.68
Persuasive	Process	0.36**	0.35	0.38	0.45

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Table 38 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Concreteness Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	−0.37**	−0.39	−0.34	−0.35
	Descriptive	0.34**	0.32	0.36	0.33
	Expository	−0.52**	−0.54	−0.50	−0.50
	Narrative	0.45**	0.43	0.47	0.46
	Persuasive	−0.37**	−0.39	−0.36	−0.40
	Process	0.24**	0.21	0.26	0.22
Compare/contrast	Descriptive	0.71**	0.69	0.73	0.69
	Expository	−0.15**	−0.17	−0.13	−0.15
	Narrative	0.82**	0.80	0.83	0.90
	Persuasive	−0.01	−0.02	0.01	−0.01
	Process	0.60**	0.58	0.62	0.59
Descriptive	Expository	−0.86**	−0.87	−0.85	−0.92
	Narrative	0.11**	0.09	0.12	0.12
	Persuasive	−0.72**	−0.73	−0.71	−0.76
	Process	−0.11**	−0.13	−0.09	−0.11
Expository	Narrative	0.97**	0.95	0.98	1.10
	Persuasive	0.14**	0.13	0.15	0.15
	Process	0.75**	0.73	0.77	0.76
Narrative	Persuasive	−0.82**	−0.83	−0.81	−0.94
	Process	−0.21**	−0.23	−0.20	−0.23
Persuasive	Process	0.61**	0.59	0.63	0.61

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 39 Mean Differences, 95% Confidence Intervals, and Effect Size Estimates Between Genres in Lexical Tightness Scores

			95% CI		
Genre A	Genre B	Mdif	*LL*	*UL*	Cohen's *d*
Cause and effect	Compare/contrast	0.50**	0.48	0.53	
	Descriptive	0.19**	0.16	0.21	
	Expository	0.02*	0.00	0.04	
	Narrative	−0.21**	−0.23	−0.19	
	Persuasive	0.08**	0.06	0.10	
	Process	0.42**	0.39	0.45	
Compare/contrast	Descriptive	−0.32**	−0.34	−0.30	
	Expository	−0.48**	−0.50	−0.46	
	Narrative	−0.72**	−0.73	−0.70	
	Persuasive	−0.43**	−0.44	−0.41	
	Process	−0.08**	−0.11	−0.06	
Descriptive	Expository	−0.16**	−0.18	−0.15	
	Narrative	−0.40**	−0.41	−0.39	
	Persuasive	−0.11**	−0.12	−0.10	
	Process	0.24**	0.21	0.26	
Expository	Narrative	−0.24**	−0.25	−0.22	
	Persuasive	0.05**	0.04	0.07	
	Process	0.24**	−0.25	−0.22	
Narrative	Persuasive	0.29**	0.28	0.30	
	Process	0.63**	0.61	0.65	
Persuasive	Process	0.35**	0.33	0.36	

Note. Post hoc comparisons are shown with Tukey–Kramer correction. CI = confidence interval. Mdif = mean difference.
*$p < .05$. **$p < .001$.

Organization. Expository essays had the highest mean organization scores, and process essays had the lowest, $F(6, 569{,}377) = 3{,}557.33$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, ps $< .05$, with the exception of the contrast between compare/contrast and process, $p = .90$. The smallest effect size was observed for the contrast between descriptive and cause and effect, whereas the largest effect sizes were observed for the contrasts between expository and process and between expository and compare/contrast (see Table 25). Overall, expository essays and narratives were high on organization, and compare/contrast and process essays were low on organization.

Sentence structure. Expository essays demonstrated the highest mean sentence structure scores, and narratives demonstrated the lowest, $F(6, 569{,}377) = 2{,}859.97$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that most contrasts were statistically significant, ps $< .05$, with the exception of the contrasts expository/persuasive, $p = .269$, and narrative/process, $p = .90$. The smallest effect size was observed for the contrast between narrative and descriptive, whereas the largest effect sizes were observed for the contrasts between expository and process and between expository and narrative (see Table 26). Overall, expository essays were associated with high sentence structure scores, and narratives were associated with low sentence structure scores.

Sentence length. Persuasive essays had the highest mean sentence length scores, followed by expository essays, cause and effect essays, compare/contrast essays, narrative/process essays, and descriptive essays, $F(6, 569{,}377) = 2{,}323.53$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, ps $< .05$. The smallest effect size was observed for the contrast between compare/contrast and narrative, $d = -.033$, whereas the largest effect size was observed for the contrast between persuasive and descriptive (see Table 27). Overall, persuasive and expository essays were longer in terms of sentence length, and descriptive essays were shorter.

Sentence complexity. Expository essays had the highest mean sentence complexity scores, and narratives had the lowest, $F(6, 569{,}377) = 1{,}835.35$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, ps $< .05$, with the exception of the contrast cause and effect/descriptive, $p = .90$. The smallest effect sizes were observed for the contrasts between descriptive and compare/contrast, cause and effect and compare/contrast, persuasive and process, and narrative and process, whereas the largest effect size was observed for the

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

contrast between expository and narrative (see Table 28). Overall, expository essays were high on sentence complexity, and narratives were low.

Grammar and usage. Overall, expository essays had the highest mean grammar and usage scores, followed by persuasive essays, cause and effect essays, narratives, descriptive essays, compare/contrast essays, and process essays, $F(6, 569{,}377) = 822.52$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that most contrasts were statistically significant, $ps < .05$, with the exception of cause and effect/compare/contrast, $p = .885$, and narrative/process, $p = .90$. The smallest effect size was observed for the contrast between persuasive and cause and effect, whereas the largest effect size was observed for the contrast between narrative and expository (see Table 29). Although there were significant genre differences, the overall pattern was one of only small genre effects for grammar.

Conventionality. Expository essays had the highest mean conventionality scores, whereas compare/contrast essays demonstrated the lowest, $F(6, 569{,}377) = 675.87$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, $ps < .05$. The smallest effect size was observed for the contrast between narrative and descriptive, whereas the largest effect size was observed for the contrast between persuasive and compare/contrast (see Table 30). Although there were significant genre differences, the overall pattern for conventionality signified only small genre effects.

Mechanics. Expository essays had the highest mean mechanics scores, whereas compare/contrast essays demonstrated the lowest, $F(6, 569{,}377) = 542.70$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with the exception of the contrast between cause and effect and narrative, $p = .817$. The smallest effect size was observed for the contrast between persuasive and cause and effect, whereas the largest effect size was observed for the contrast between expository and compare/contrast (see Table 31). Although there were significant genre differences, the overall pattern was one of only small genre effects for mechanics.

Narrativity. Narratives had the highest narrativity scores, whereas persuasive essays had the lowest, $F(6, 569{,}377) = 49{,}380.56$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with the exception of the contrast between compare/contrast and process. The smallest effect size was observed for the contrast between descriptive and cause and effect, whereas the largest effect size was observed for the contrast between narrative and persuasive (see Table 32). Overall, narratives were high on the narrativity trait, whereas expository, compare/contrast, cause and effect, and persuasive essays were low on the narrativity trait.

Contextualization. Narratives had the highest contextualization scores, whereas persuasive essays had the lowest, $F(6, 569{,}377) = 51{,}818.61$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with the exception of the contrast between compare/contrast and process, $p = .256$. The smallest effect size was observed for the contrast between descriptive and cause and effect, whereas the largest effect size was observed for the contrast between narrative and persuasive (see Table 33). Overall, narratives were high on the contextualization trait, whereas compare/contrast, process, and persuasive essays were low.

Dialog. Narratives had the highest dialog scores, whereas persuasive essays had the lowest, $F(6, 569{,}377) = 17{,}653.50$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that most contrasts were statistically significant, $ps < .05$, with the exception of the contrasts between compare/contrast and process, $p = .90$, and between descriptive and expository, $p = .069$. The smallest effect size was observed for the contrast between expository and compare/contrast, whereas the largest effect size was observed for the contrast between narrative and persuasive (see Table 34). Overall, narratives were high on the dialog trait, whereas compare/contrast, process, expository, and persuasive essays were low.

Cohesion. Compare/contrast essays showed the highest mean cohesion scores, whereas narratives showed the lowest, $F(6, 569{,}377) = 6{,}596.90$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with the exception of the contrast between compare/contrast and process, $p = .281$. The smallest effect size was observed for the contrast between expository and cause and effect, whereas the largest effect size was observed for the contrast between narrative and process (see Table 35). Overall, compare/contrast, process, and persuasive essays were high on cohesion, and narratives were low.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Stance Taking. Persuasive essays had the highest mean stance taking scores, whereas narratives had the lowest, $F(6, 569{,}377) = 32{,}125.40$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, $ps < .05$. The smallest effect sizes were observed for the contrasts between compare/contrast and process and between persuasive and cause and effect, whereas the largest effect size was observed for the contrast between narrative and persuasive (see Table 36). Overall, persuasive, cause and effect, and expository essays were on high on the stance taking trait, whereas narratives and descriptive essays were low.

Interactivity. Narratives had the highest mean interactivity scores, whereas persuasive essays demonstrated the lowest, $F(6, 569{,}377) = 13{,}096.01$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, $ps < .05$. The smallest effect size was observed for the contrast between descriptive and compare/contrast, whereas the largest effect size was observed for the contrast between narrative and persuasive (see Table 37). Overall, narratives were high and persuasive essays were low on the interactivity trait.

Concreteness. Narratives demonstrated the highest mean concreteness scores, and expository essays demonstrated the lowest, $F(6, 569{,}377) = 15{,}820.79$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that contrasts were statistically significant, $ps < .05$, with the exception of the contrast between compare/contrast and persuasive, $p = .701$. The smallest effect size was observed for the contrast between descriptive and process, whereas the largest effect size was observed for the contrast between expository and narrative (see Table 38). Overall, narrative and descriptive essays were high on concreteness, whereas compare/contrast, persuasive, and expository essays were low on concreteness.

Lexical tightness. Compare/contrast essays had the highest mean lexical tightness scores, whereas narratives had the lowest, $F(6, 569{,}377) = 3{,}463.03$, $p < .001$ (see Table 21). Post hoc comparisons with a Tukey–Kramer correction showed that all contrasts were statistically significant, $ps < .05$. The smallest effect size was observed for the contrast between expository and cause and effect, whereas the largest effect size was observed for the contrast between narrative and compare/contrast (see Table 39). Overall, compare/contrast and process essays were high on the lexical tightness dimension, and narrative essays were low.

Research Question 4: Discussion

As these results indicate, there are significant, and sometimes quite large, trait differences among the seven Criterion topic library genres. First, narratives are quite distinct. Naturally, they are high on narrativity, hence they are high on the traits of contextualization and dialog but low on cohesion. Narratives are also high on interactivity and organization. Narratives tend to demonstrate lower stance taking, sentence structure (specifically, on sentence complexity), and formality. The lower formality trait further rationalizes narratives' lower vocabulary length but higher vocabulary frequency features.

Persuasive essays fall almost diametrically opposite to narratives. Persuasives were low on narrativity, hence low on contextualization and dialog and high on cohesion. Persuasives were also low on interactivity, high on stance taking, high on formality (hence high on vocabulary length and low on vocabulary frequency), and high on sentence length, but also high on organization. Expository essays fall close on many dimensions to persuasive essays. Expository essays were low on narrativity (specifically, low on dialog), but they were high on stance taking, formality (specifically, high on vocabulary length), sentence structure (hence high on sentence length and sentence complexity), and organization.

Compare/contrast essays were low on narrativity (hence low on contextualization and dialog), organization, and concreteness; however, they were high on lexical tightness. Process essays were similar to compare/contrast essays, as they were high on lexical tightness and low on organization. However, compare/contrast essays were low on formality (hence on vocabulary length), dialog, and concreteness, yet they were high on cohesion. Cause and effect essays, like expository and persuasive essays, were low on narrativity (hence low on contextualization) and high on stance taking, but unlike expository and persuasive essays, they were high on concreteness. Finally, descriptive essays were low on stance taking and sentence length but high on concreteness and vocabulary frequency.

Overall, these patterns make sense given the focus of each genre. Critically, there were significant effects for nearly every trait–genre combination, and the effect sizes were consistently small only for the conventionality, mechanics, and grammar traits. These results confirm the hypothesis that motivated the trait model in the first place, namely, the conviction that genre differences matter for AWE. Each genre has a unique trait profile. This finding implies that the same

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 40 Loadings on Regression Terms for Standard Error of Measurement Predicting Traits From Private School Status

Trait	Estimate	Estimated SD	*SE*	*z*-value	*p*-value
Formality	0.007	0.004	0.003	2.26	0.024
Grammar and usage	0.197	0.142	0.003	64.54	<0.001
Concreteness	0.069	0.034	0.004	18.82	<0.001
Sentence complexity	0.068	0.058	0.002	29.59	<0.001
Vocabulary length	0.053	0.039	0.003	19.19	<0.001
Mechanics	0.008	0.007	0.003	2.47	0.013
Interactivity	0.004	0.003	0.003	1.61	0.108
Organization	−0.009	−0.004	0.002	−4.04	<0.001
Contextualization	−0.030	−0.026	0.003	−8.81	<0.001
Vocabulary frequency	−0.060	−0.032	0.003	−17.98	< 0.001
Dialog	−0.066	−0.111	0.003	−20.70	<0.001
Cohesion	−0.074	−0.032	0.004	−19.44	<0.001
Stance taking	−0.089	−0.038	0.004	−23.62	<0.001
Sentence length	−0.094	−0.090	0.002	−47.13	<0.001

student may show higher levels on any one trait (e.g., organization) or lower levels on another trait (e.g., stance taking) entirely owing to the nature of the task.

Research Question 5: How Are the Traits Affected by Demographic Variables?

Research Question 5: Results

We entered the demographic variables as additional regression coefficients into the structural equation (SEM) model we originally used to estimate trait scores. When we did so, fit increased slightly (CFI = .933; RMSEA = .0460). Most of the demographic variables were significant predictors of the trait scores, but the weights were generally very small.

Table 40 shows the estimates for private/public school status as a predictor of each of the subordinate trait scores.[8] Only grammar and usage demonstrated a loading greater than .10 (+.197), indicating that students at private schools were somewhat more likely to avoid grammar and usage errors than students at public schools. None of the other demographic variables (school location, school full-time equivalent staff (FTE), student/teacher ratio, or race/ethnicity [White, African American, Hispanic, Asian, or Native American]) produced loadings on any trait score greater than .10.

We also ran a second SEM in which we restricted the data to the public schools in the set, because we had information about gender and free or reduced-price school lunch status only for these schools. Adding these variables to the model had little appreciable impact. The fit was slightly better for this restricted data set (CFI = .936; RMSEA = .0442), but although gender and free/reduced-price lunch status were significant predictors, the estimates we obtained for these variables were also very close to zero, $p < .01$.

Research Question 5: Discussion

The demographic effects we observed were, for the most part, extremely small. The only exception was the effect on the grammar and usage trait associated with private school status. It is difficult to be certain why schools using Criterion would not have displayed stronger demographic effects. One possibility is that demographic effects were present at the individual level but not so clearly evident when the data were aggregated at the school level. Another possibility is that when students were writing classroom essays, they may have been more motivated and/or had more time to complete the task compared to when students take on-demand writing assessments, which typically have a strict time limit. This last possibility would be very interesting if it were true, as it might suggest that on-demand writing assessments underestimate the writing achievement of students from underserved groups. At this point, the data are not sufficient to distinguish among the various interpretations. However, the findings raise interesting questions for future research. Are there interactions between on-demand versus classroom writing and demographic variables? Under what circumstances are the effects of demographic variables muted, as they appear to be in this data set?

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 41 Correlations Between Trait Scores and Imputed Student Grade Level

Trait	Correlation with grade level	
	All prompts	Topic library prompts
Formality	.397	.468
Vocabulary length	.330	.410
Organization	.263	.254
Interactivity	−.221	−.284
Concreteness	−.241	−.336
Sentence complexity	.194	.241
Sentence structure	.192	.238
Vocabulary frequency	−.155	−.158
Sentence length	.144	.164
Grammar and usage	.086	.154
Stance taking	.086	.131
Conventionality	.081	.097
Narrativity	−.073	−.211
Dialog	−.068	−.193
Contextualization	−.053	−.179
Mechanics	.049	.059
Cohesion	−.040	.031

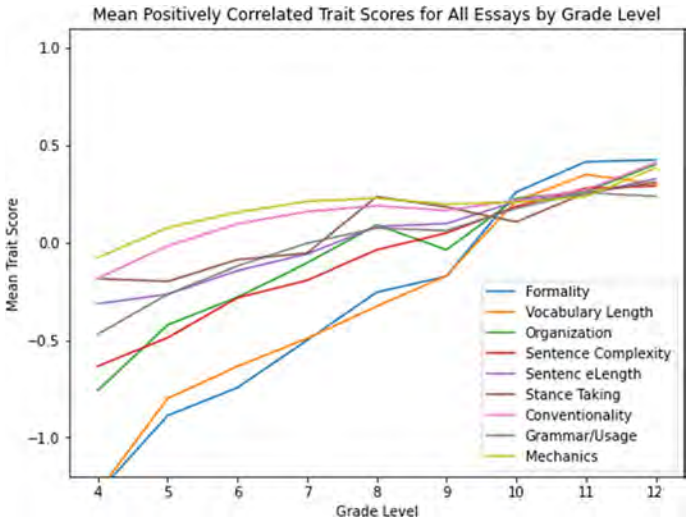

Figure 12 Trend lines for positively correlated traits by imputed grade level.

Research Question 6a: Can the Model Be Used to Measure Growth in Specific Traits Within and Across Grades?

Research Question 6a: Results

We examined the correlation between imputed student grade levels and trait scores for the entire data set. The results are shown in Table 41, by descending absolute magnitude of the correlation. We examine both the correlation with all essays and the correlation with topic library prompts, where the tasks are somewhat more comparable across grades.

Note that several traits strongly associated with genre differences (e.g., stance taking, contextualization, cohesion, lexical tightness) are among the traits with the lowest correlations with grade level. Interestingly, the grade-level trends appear to be stronger for topic library prompts than for the entire corpus of writing tasks.

If we plot the positively and negatively correlated trait means separately by grade level, as shown in Figures 12 and 13, we observe that the increasing or decreasing trends appear to be quite smooth, with only small perturbations from a linear pattern of increase, decrease, or stability.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

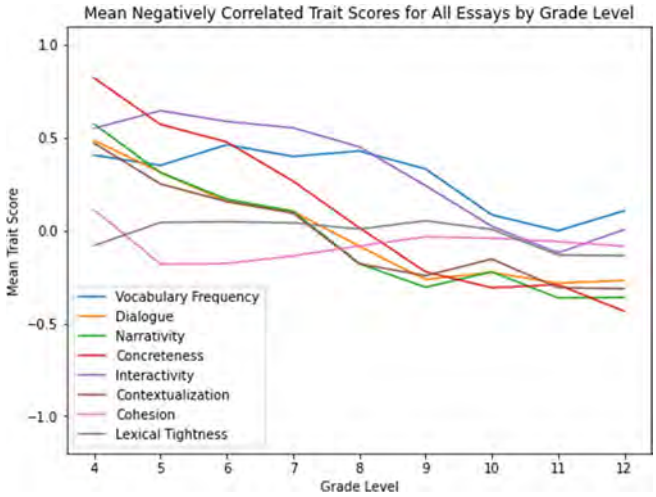

Figure 13 Trend lines for negatively correlated traits by imputed grade level.

Table 42 Frequency of Criterion Topic Library Submissions by Genre and Grade

Grade	Cause and effect	Compare/contrast	Descriptive	Expository	Narrative	Persuasive	Process
4	1,371	796	1,315	11,440	5,383	5,760	1,235
5	3,365	5,415	7.980	18,592	13,171	13,944	3,763
6	8,023	15,431	20,399	72,445	32,381	67,657	11,841
7	7,875	11,013	19,618	108,626	26,008	74,911	11,419
8	4,523	5,612	15,243	130,821	16,929	86,215	2,322
9	838	323	12,981	89,536	9,714	46,591	123
10	127	210	12,676	86,005	12,137	75,667	83
11	73	122	6,675	69,424	9,419	52,700	18
12	15	31	1,927	39,056	3,695	30,031	10

A somewhat more complex set of issues arises when we examine the relationship between grade level and genre. Table 42 tabulates the number of submissions per grade by genre for Criterion topic library prompts. Four of the genres—descriptive, expository, narrative, and persuasive—are well represented across the grade range. The other three genres—cause and effect, compare/contrast, and process—appear to have been predominantly assigned in middle school, as 92% of the cause and effect essays, 82% of the compare/contrast essays, and 83% of the process essays were assigned in sixth, seventh, and eighth grades. The sample sizes for elementary and high school are quite low. As a result, these genres are not useful for examining grade-level trends. Thus we restrict our attention to the four genres that are well represented across the grade span.

Figures 14–18 show the grade-level means by genre for the five traits with the strongest grade-level trends (formality, vocabulary length, organization, concreteness, and sentence complexity). As examination of Figures 14–18 indicates, the trend lines for the four genres are essentially parallel, with only minor perturbations for particular genres at particular grade levels, which might reflect smaller sample sizes in some grades or the mix of specific assignments and topics chosen by teachers at that grade. Of course, the trend lines are offset up or down, based on genre.

Figures 19–23 show the means for the same five traits by month of the school year, rather than by year. We observed somewhat noisier patterns, with mostly small jumps up and down within the main trend line. Larger deviations from the trend line are associated mostly with the description genre, for which the samples by month are sometimes missing or very small.

Research Question 6a: Discussion

The results reported here support a developmental pattern in which traits like organization, formality, vocabulary length, sentence complexity, and concreteness systematically increase (or, in the case of concreteness, decrease) by grade level.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

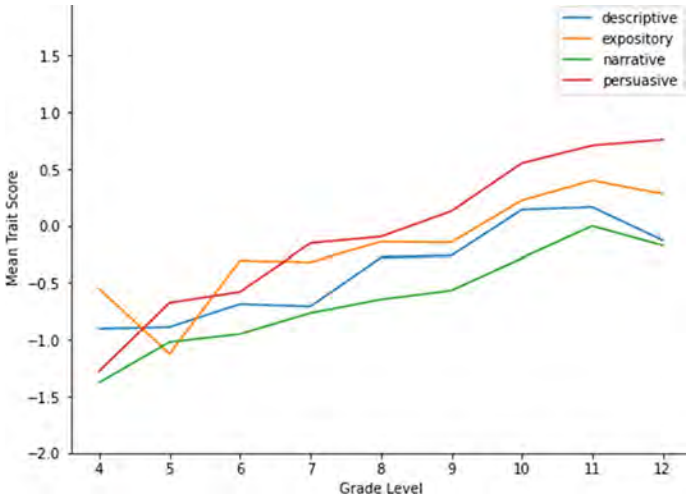

Figure 14 Mean formality scores by genre and grade level.

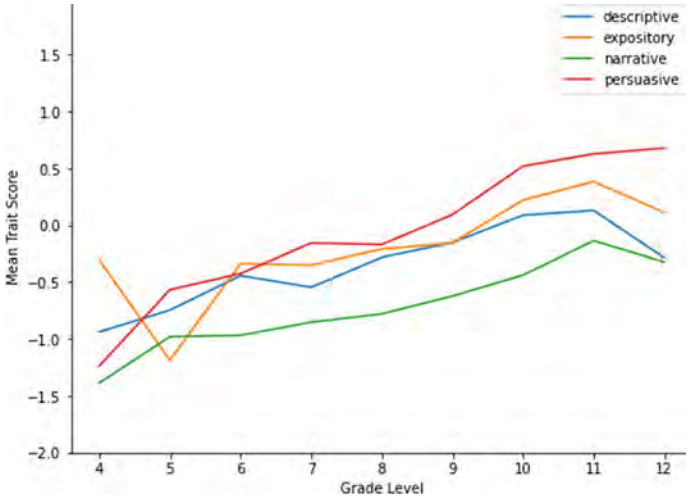

Figure 15 Mean vocabulary length scores by genre and grade level.

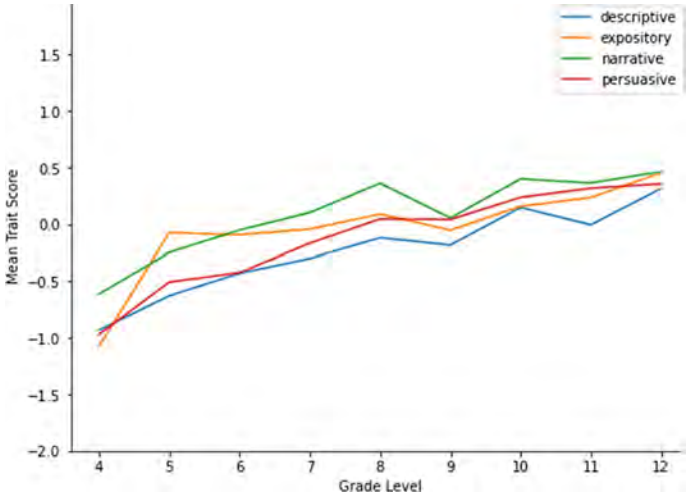

Figure 16 Mean organization scores by genre and grade level.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

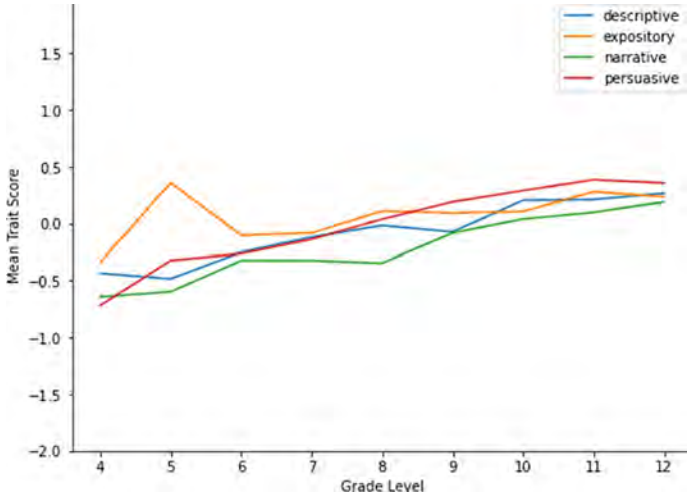

Figure 17 Mean sentence complexity scores by genre and grade level.

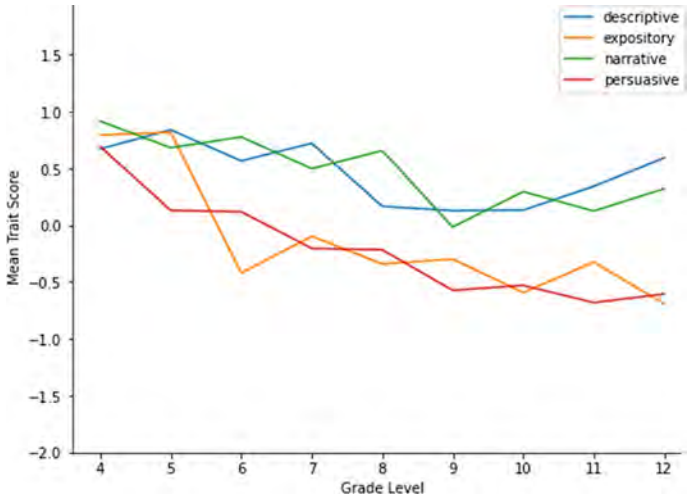

Figure 18 Mean concreteness scores by genre and grade level.

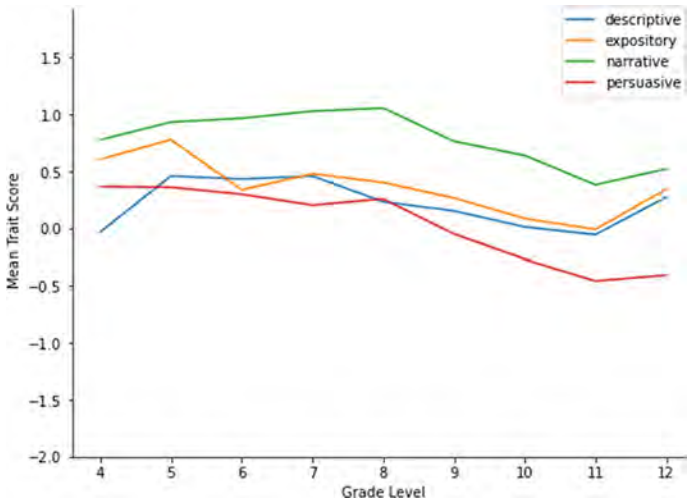

Figure 19 Mean interactivity scores by genre and month in school year.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

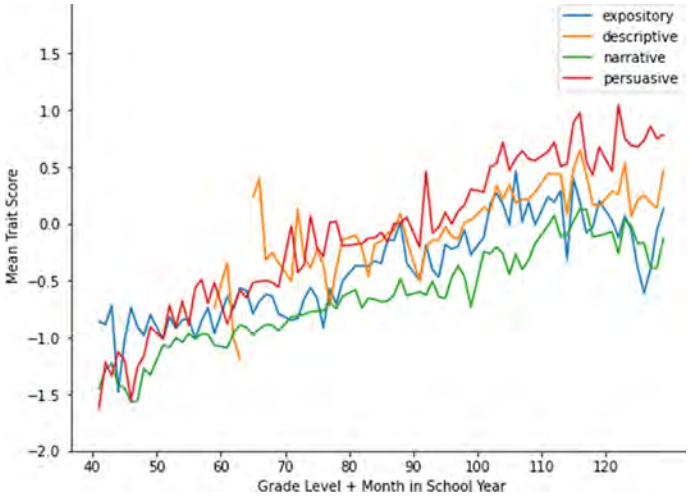

Figure 20 Mean vocabulary length scores by genre and month in school year.

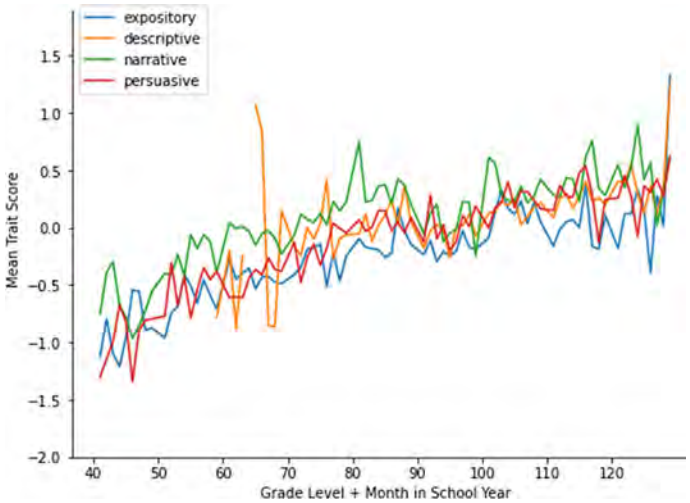

Figure 21 Mean organization scores by genre and month in school year.

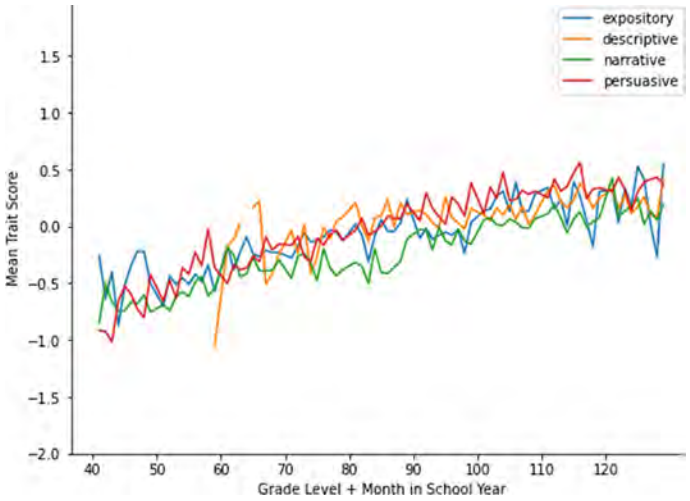

Figure 22 Mean sentence complexity scores by genre and month in school year.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

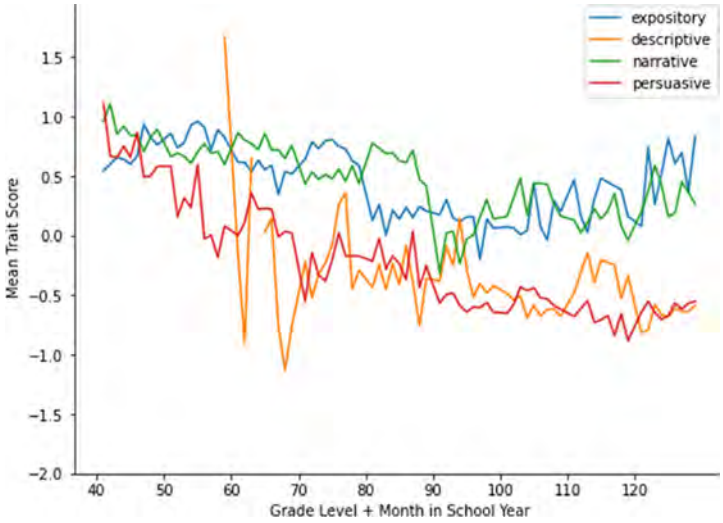

Figure 23 Mean concreteness scores by genre and month in school year.

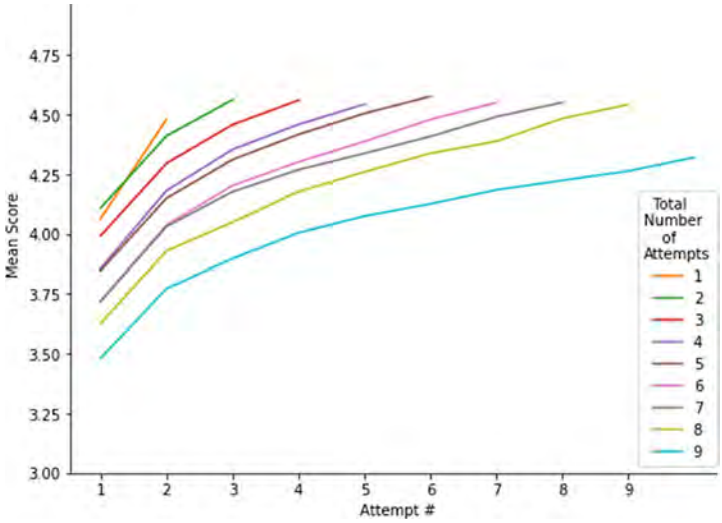

Figure 24 Changes in mean e-rater score after revision by total number of drafts.

The data appear likely to fit a simple linear model that specifies separate effects by trait for age and genre. This means that it should be possible to use scores on writing tasks during the school year to track progress, at least at a group level, and to identify students whose performance on specific writing assignments falls significantly above or below their expected trend lines. Because we do not know what instruction students were given at any specific school, the Criterion data do not provide direct evidence that the traits can be used to measure growth after an intervention; however, it is likely that the measures are sensitive and reliable enough to be used for that purpose.

Research Question 6b: Can the Model Be Used to Evaluate Changes in Specific Traits After Revision?

Research Question 6b: Results

We examined mean trait scores for successive attempts of the same essay to investigate the extent to which trait scores change as a result of revision. Because essays that received more revisions tended to have lower e-rater scores, we plotted separate means based on the number of drafts an essay went through. The results for e-rater scores are shown in Figure 24.

Mean e-rater scores increased with each successive draft, though the magnitude of the score increase dropped as the number of drafts increased.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

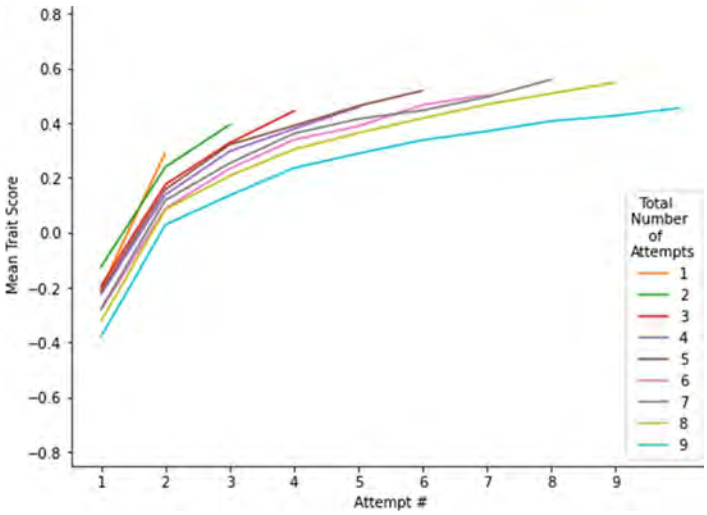

Figure 25 Changes in mean grammar and usage score after revision by total number of drafts.

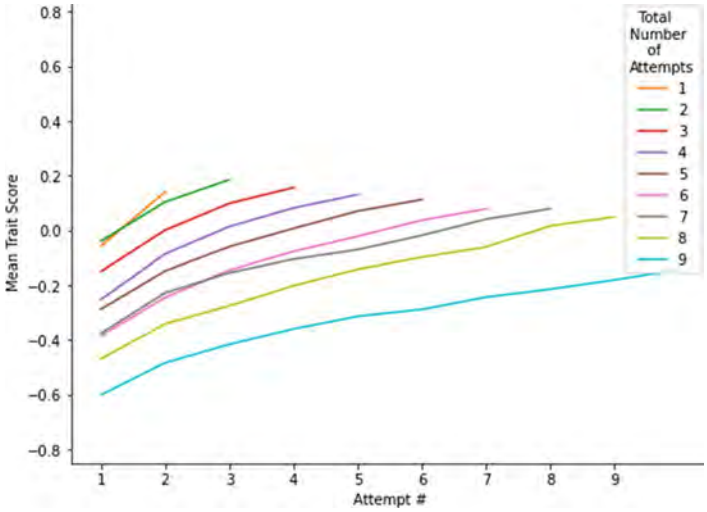

Figure 26 Changes in mean organization score after revision by total number of drafts.

Several traits showed similar patterns, though the magnitude of the effect varied (see Figures 25 – 31). The largest effects were for grammar and usage, with somewhat smaller effects for organization and formality, followed by sentence complexity, lexical tightness, vocabulary length, and formality. The effects for grammar and usage were relatively large on the first revision (about half a point on the trait score scale). For the other traits that showed increases, the effects were smaller. Some of the traits showed little mean improvement after the first revision — most notably, conventionality. In addition, the traits most strongly associated with genre (e.g., narrativity, stance taking) showed little change between revisions.

Research Question 6b: Discussion

These results indicate that the trait model is sensitive enough to detect changes between drafts, and on average, writing quality increased between successive drafts. The largest effects were for grammar and usage, organization, formality, sentence complexity, lexical tightness, and conventionality. However, mean score increases were smaller for second and later revisions. It is interesting to note that Criterion's automated feedback focuses on grammar, usage, mechanics, style (including sentence length and repeated words), and essay organization. Thus the relatively large effects we see for grammar and usage and organization may be due to students addressing the feedback that Criterion explicitly provides. Similarly, the decrease in the pace of improvement after the first revision may be related to the fact that students often use Criterion on

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

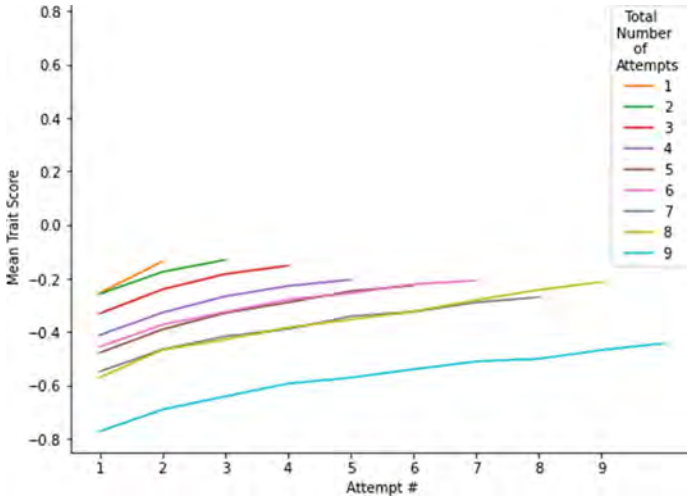

Figure 27 Changes in mean formality score after revision by total number of drafts.

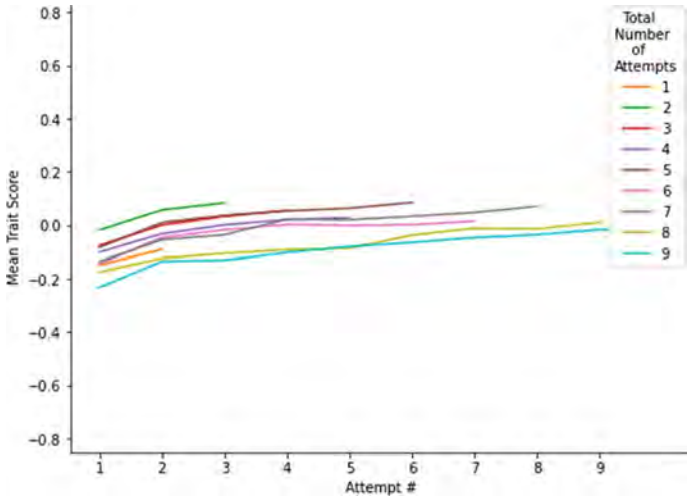

Figure 28 Changes in mean sentence complexity score after revision by total number of drafts.

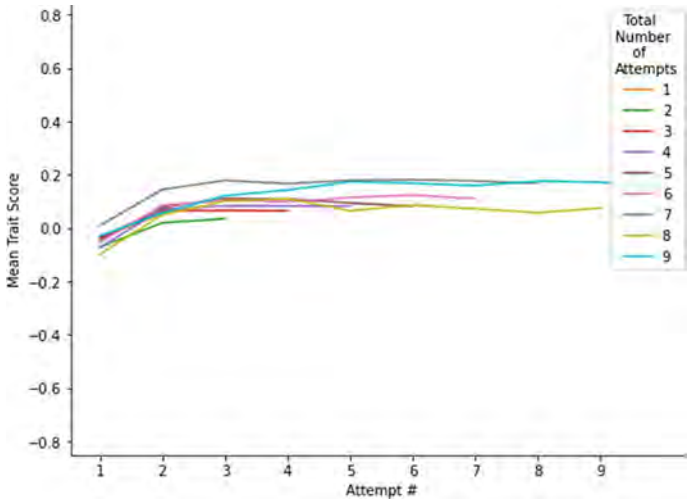

Figure 29 Changes in mean lexical tightness score after revision by total number of drafts.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

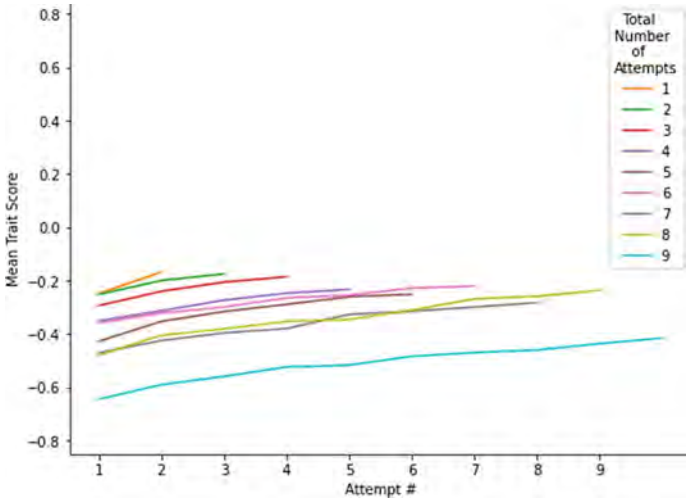

Figure 30 Changes in mean vocabulary length score after revision by total number of drafts.

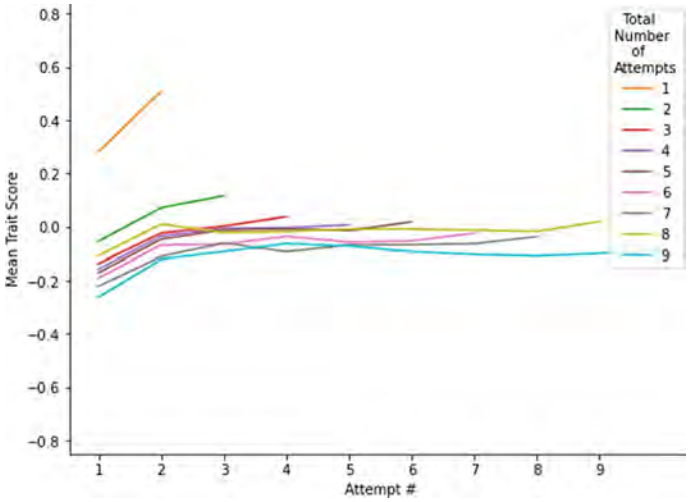

Figure 31 Changes in mean conventionality score after revision by total number of drafts.

their own as a tool before they submit their essays to teachers. In our database of Criterion assignments, we have frequently observed teachers telling students not to submit the essay to them until they achieve a minimum e-rater score.

Research Question 6c: Can the Model Be Used to Assess Overall Writing Quality?

Research Question 6c: Results

Human-Scored Essays Written to Criterion Prompts

The correlations between trait scores and human scores for the human-scored Criterion data set are shown in Table 43. Of all the traits, organization had the highest correlation with human scores. Formality came next, followed by conventionality, mechanics, grammar and usage, vocabulary length, lexical tightness, vocabulary frequency, and sentence complexity. All other absolute correlations were small.

When we trained automated scoring models using ordinary least squares regression and entered all predictors simultaneously, 56 of the 71 trait-based models (78.9%) had the highest adjusted R^2 compared to the e-rater feature-based models. In addition, 41 of the 71 trait-based models (57.7%) had the highest quadratic weighted kappa on the test set (see Table 44). On average, the models that used only standard e-rater features had an adjusted R^2 of .730 and achieved a

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 43 Correlations Between Traits and Human Essay Scores for Human-Scored Essays Written to Criterion Prompts

Trait	Correlation
Formality	.44
Vocabulary length	.32
Vocabulary frequency	−.25
Organization	.66
Sentence structure	.21
Sentence length	.03
Sentence complexity	.25
Grammar and usage	.35
Conventionality	.39
Mechanics	.36
Narrativity	.16
Contextualization	.12
Dialog	.13
Cohesion	−.02
Stance taking	−.11
Interactivity	−.20
Concreteness	.07
Lexical tightness[a]	.25

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

quadratic weighted kappa of .785 on the test set. The models that supplemented e-rater features by adding the grammaticality and lexical tightness features had an average adjusted R^2 of .742 and achieved an average quadratic weighted kappa of .793 on the test set. The models that employed the trait model feature set had an average R^2 of .751 and achieved an average quadratic weighted kappa of .798 on the test set. Overall, the trait-based model (i.e., the model created by supplementing the trait scores with e-rater's LOGDTA and LOGDTU features to account for the effects of overall composition fluency) appears to have yielded more accurate predictions of human holistic scores on these prompts than either of the e-rater feature-based models we tested.[9]

Human-Scored Essays Written to CBAL Prompts

Somewhat similar results were evident when we examined the essay data from the CBAL Reading and Writing assessments. The organization trait score had the highest correlation with human scores, followed by lexical tightness, conventionality, grammar, sentence complexity, concreteness, contextualization, vocabulary length, and formality. The other traits (e.g., stance taking, cohesion) had relatively small correlations with human scores (see Table 45).

When we built automated scoring models for the CBAL prompts, we obtained the pattern of results shown in Table 46. The trait-based model had the highest adjusted R^2 in 100% of the models and achieved the highest quadratic weighted kappa on the test set in 13 of the 17 models (76.5%). The average adjusted R^2 for the models based on standard e-rater features was .551, with a quadratic weighted kappa of .630 on the test set. The average adjusted R^2 for the extended e-rater models that included the grammaticality and lexical tightness features was .552, with an average quadratic weighted kappa of .621 on the test set. The average adjusted R^2 for the trait-based model was .582, with an average quadratic weighted kappa of .666 on the test set. Once again, the trait-based model achieved the highest average results (see Table 46).

The CBAL data set was one of the few human-scored data sets for which we had genre information for all responses. Three genres (literary analysis, policy argument, and cost–benefit analysis essays) were used across multiple prompts in multiple CBAL studies, enabling us to build genre-specific scoring models. To obtain reliable trait loadings for these models, we excluded the superordinate (and therefore collinear) traits formality, narrativity, sentence structure, and conventionality from the regression, leaving only the subordinate traits. Table 47 shows the pattern of weights we obtained in the resulting models for each genre.[10]

Certain traits make distinct contributions to the three models. In the literary analysis model, $F(16, 1{,}787) = 79.66$, $p < .001$, $R^2 = .416$, lexical tightness had a weight of .267, interactivity a weight of −.273, and dialog a weight of .187. These weights contrast strongly with the same features' weights in the other two models. The policy argument model,

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 44 Performance of e-rater and Trait Models on Human-Scored Essays Written to Criterion Prompts

			Standard e-rater features		e-rater plus supplemental features		Trait model	
Prompt ID	*N* train	*N* test	Adj. R^2	QWK	Adj. R^2	QWK	Adj. R^2	QWK
ECRT1188	499	496	.78	.777	.80	.791	.80	.794
ECPT0000	2,225	2,224	.77	.795	.78	.795	.79	.800
ECRT1194	482	474	.80	.799	.81	.801	.81	.793
ECRT1154	233	233	.82	.848	.84	.848	.85	.836
ECRT1213	228	231	.83	.869	.84	.867	.85	.877
ECRT1101	234	232	.80	.858	.82	.880	.85	.879
ECXT0000	3,212	3,204	.77	.803	.78	.807	.79	.809
ECRT1175	234	235	.79	.844	.80	.851	.82	.867
ECRT0225	233	231	.85	.890	.86	.889	.86	.901
ECRT1172	487	487	.75	.787	.75	.791	.77	.803
ECRT0342	233	235	.80	.848	.81	.868	.81	.866
ECRT1219	484	476	.81	.845	.82	.834	.82	.831
ECRT1185	230	229	.83	.848	.84	.870	.87	.879
ECRT2215	233	231	.88	.843	.90	.857	.89	.859
ECRT1168	231	230	.80	.836	.83	.853	.83	.844
ECRT1246	229	228	.83	.872	.84	.865	.85	.889
ECRT1161	234	234	.79	.866	.81	.864	.83	.878
ECRT1223	231	230	.86	.856	.88	.866	.89	.872
ECRT0230	232	230	.84	.870	.86	.868	.86	.899
ECRT1190	233	233	.79	.838	.80	.848	.81	.876
ECRT1162	235	235	.81	.834	.82	.856	.83	.831
ECRT1114	232	232	.87	.872	.89	.882	.89	.887
ECRT0349	234	235	.80	.799	.81	.819	.82	.833
ECRT1234	228	226	.85	.837	.87	.856	.88	.858
ECRT1170	234	233	.84	.843	.86	.850	.86	.844
ECRT0279	231	232	.87	.848	.87	.879	.88	.866
ECRT1100	231	230	.83	.847	.84	.872	.86	.865
ECRT1215	221	228	.84	.864	.85	.866	.84	.884
ECRP0014	282	281	.58	.705	.58	.688	.56	.688
ECRP0044	500	498	.70	.708	.70	.704	.70	.724
ECRR0781	232	231	.42	.569	.43	.628	.48	.620
ECRR0612	394	220	.79	.807	.80	.810	.90	.816
ECRR0349	227	228	.69	.718	.69	.725	.70	.741
ECRR0631	256	256	.83	.818	.83	.817	.83	.790
ECRR0596	184	187	.77	.815	.77	.818	.80	.819
ECRR2342	205	209	.78	.835	.79	.844	.80	.847
ECRR0111	183	182	.63	.739	.64	.739	.64	.773
ECRR0785	242	244	.57	.649	.58	.650	.58	.647
ECRR1536	193	196	.79	.834	.80	.839	.81	.836
ECRR0570	232	230	.71	.827	.71	.825	.73	.814
ECRR1547	273	274	.83	.852	.83	.851	.85	.849
ECRR0115	248	245	.56	.780	.58	.775	.59	.757
ECRR1688	190	195	.53	.627	.53	.605	.55	.667
ECRR0634	266	263	.79	.841	.80	.833	.81	.844
ECRE000G	385	290	.48	.584	.50	.614	.49	.607
ECRE000H	390	240	.36	.626	.37	.635	.41	.613
ECRE000I	490	437	.56	.660	.58	.688	.60	.681
ECRE000E	215	214	.64	.680	.65	.694	.64	.685
ECRE000F	496	494	.56	.712	.57	.713	.58	.740
ECRE000A	209	209	.61	.686	.62	.666	.64	.672
ECRE000D	210	206	.68	.703	.68	.689	.68	.683
ECRE000B	196	202	.68	.769	.69	.775	.69	.748
ECRE000C	199	198	.65	.704	.70	.740	.68	.758
ECRD6200	443	444	.85	.816	.85	.817	.85	.821
ECRD6400	1,011	1,013	.69	.742	.71	.754	.72	.779
ECRD6700	361	350	.75	.776	.76	.778	.79	.798

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 44 Continued

Prompt ID	*N* train	*N* test	Standard e-rater features		e-rater plus supplemental features		Trait model	
			Adj. R^2	QWK	Adj. R^2	QWK	Adj. R^2	QWK
ECRD6600	453	464	.6	.707	.63	.707	.66	.738
ECRD6100	480	490	.79	.792	.79	.799	.80	.799
ECRD6000	669	667	.76	.796	.77	.807	.77	.813
ECRD6900	356	360	.74	.778	.75	.770	.77	.800
ECRD6800	506	498	.74	.775	.75	.783	.78	.801
ECRN0008	395	270	.66	.792	.68	.802	.69	.816
ECRN0012	267	264	.77	.806	.77	.809	.80	.823
ECRN004B	213	213	.70	.809	.72	.842	.73	.841
ECRN008B	169	169	.67	.772	.68	.799	.69	.787
ECRN004C	176	183	.63	.748	.64	.755	.60	.744
ECRN004A	199	200	.75	.782	.76	.800	.77	.782
ECRN012A	176	177	.69	.786	.71	.825	.70	.817
ECRN012B	162	162	.70	.805	.72	.817	.77	.842
ECRN008A	172	173	.70	.727	.73	.774	.71	.809
ECRC6R00	3,416	3,427	.58	.690	.60	.696	.61	.702
Average			*.73*	*.785*	*.74*	*.793*	*.75*	*.798*

Note. QWK = quadratic weighted kappa.

Table 45 Correlations Between Traits and Human Essay Scores for Human-Scored Essays Written to CBAL Prompts

Trait	Correlation
Formality	.35
Vocabulary length	.19
Vocabulary frequency	.04
Organization	.61
Sentence structure	.07
Sentence length	−.03
Sentence complexity	.17
Grammar and usage	.17
Conventionality	.28
Mechanics	.22
Narrativity	.32
Contextualization	.24
Dialog	.28
Cohesion	.13
Stance taking	−.06
Interactivity	.11
Concreteness	.24
Lexical tightness[a]	.44

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

$F(16, 7{,}535) = 290.1$, $p < .001$, $R^2 = .38$, and the cost–benefit analysis model, $F(16, 3{,}008) = 336.8$, $p < .001$, $R^2 = .642$, had relatively similar weights, reflecting the fact that both genres are persuasive in purpose, which may be why both models displayed significant negative weights on the contextualization trait. However, cohesion, lexical tightness, and interactivity were significant predictors in the policy argument model but not in the cost–benefit model. Furthermore, stance taking was a significant predictor in the cost–benefit model but not in the policy argument model. These results support the hypothesis that that genre-differentiating traits like interactivity, contextualization, dialog, cohesion, and lexical tightness may have predictive value in genre-specific essay scoring models, even if they do not have much predictive value across genres.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 46 Performance of e-rater and Trait Models on Human-Scored Essays Written to CBAL Prompts

			Standard e-rater features		Extended e-rater model		Trait model	
Prompt ID	*N* train	*N* test	Adj. R^2	QWK	Adj. R^2	QWK	Adj. R^2	QWK
House on Mango Street	1,804	435	.40	.602	.40	.607	.41	.610
Service Learning	1,389	327	.63	.706	.63	.703	.65	.751
Ban Ads	3,952	991	.25	.375	.25	.373	.25	.395
Invasive Species B	52	11	.28	.199	.29	.014	.43	.619
Cash for Grades	2,186	533	.64	.770	.65	.775	.66	.784
Generous Gift	827	221	.64	.777	.65	.779	.66	.799
Culture Fair	809	230	.62	.692	.63	.692	.64	.729
Social Networking	752	198	.69	.806	.69	.803	.70	.817
Fender Bender	95	18	.55	.649	.54	.620	.61	.723
Dolphin Intelligence	239	42	.80	.840	.76	.845	.79	.867
Organic Farming	316	83	.47	.581	.47	.606	.47	.580
Task 1: Teaching Cursive Writing	111	32	.71	.652	.71	.669	.71	.652
Task 2: Eliminating Paper Mail	101	38	.37	.645	.35	.645	.40	.646
Task 2: Human Space Exploration	116	22	.31	.249	.30	.353	.37	.443
Task 3: Foreign Language Requirement	114	27	.76	.760	.76	.746	.77	.706
Task 3: Human Space Exploration	106	35	.68	.729	.68	.729	.70	.601
Task 1: Foreign Language Requirement	114	28	.63	.686	.63	.604	.68	.606
Average			*.55*	*.630*	*.55*	*.621*	*.58*	*.664*

Note. QWK = quadratic weighted kappa.

Table 47 Trait Loadings for Three CBAL Genres

Subordinate trait	Literary analysis	Policy argument	Cost–benefit analysis
Organization	1.07**	1.79**	1.56**
Residual LOGDTU	0.15	0.44**	0.14*
LOGDTA	0.66**	0.38**	0.35**
Vocabulary length	0.33**	0.12**	0.21**
Vocabulary frequency	−0.01	−0.09*	−0.08*
Sentence length	−0.08	−0.09	−0.00
Sentence complexity	0.09	0.02	0.02
Grammar and usage	0.14*	0.14**	0.19**
Mechanics	0.17*	−0.04	0.10**
Contextualization	−0.02	−0.19**	−0.23**
Dialog	0.19**	−0.05	−0.06*
Cohesion	0.07	−0.06*	0.01
Stance taking	0.08	−0.02	0.10**
Interactivity	−0.27**	0.13**	0.07
Concreteness	−0.06	−0.09*	−0.10**
Lexical tightness[a]	0.27**	−0.07*	0.00

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them. $^{*}p < .05$. $^{**}p < .01$.

Human-Scored Essays Written to HiSET Prompts

Results for the HiSET prompts were fairly similar to what we have observed thus far. The organization trait had the highest correlation with human scores, followed by the formality factor score, the conventionality factor score, the grammar and usage factor score, the vocabulary length factor score, and the sentence complexity factor score. All other traits had small correlations (see Table 48).

In this data set, the trait-based model had the highest adjusted R^2 on all 10 data sets. On the test set, it had the highest quadratic weighted kappa for 6 of the 10 prompts (see Table 49). The models that used only standard e-rater features had an average adjusted R^2 of .60 and achieved an average quadratic weighted kappa of .655 on the test set. The models that supplemented standard e-rater features by including the grammaticality, cohesion, and lexical tightness features had an

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 48 Correlation of Trait Scores to Human Holistic Essay Scores for HiSET Prompts

Trait	Correlation
Formality	.38
Vocabulary length	.19
Vocabulary frequency	−.07
Organization	.74
Sentence structure	.14
Sentence length	.04
Sentence complexity	.18
Grammar and usage	.22
Conventionality	.28
Mechanics	.24
Narrativity	.21
Contextualization	.15
Dialog	.16
Cohesion	−.03
Stance taking	−.07
Interactivity	.03
Concreteness	.09
Lexical tightness[a]	.10

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

Table 49 Performance of e-rater and Trait Models for Essays Written to HiSET Prompts

			Standard e-rater features		Extended e-rater model		Trait model	
Prompt ID	*N* train	*N* test	R^2	QWK	R^2	QWK	R^2	QWK
VH581470	1,277	346	.62	.682	.63	.670	.64	.666
VH581467	1,499	375	.60	.626	.62	.621	.62	.590
VH337030	3,556	924	.62	.676	.64	.675	.65	.686
VH337290	3,518	895	.62	.661	.63	.678	.64	.674
VH730325	3,067	772	.67	.700	.68	.714	.69	.717
VH581474	1,393	383	.59	.645	.60	.653	.61	.677
VH730335	3,045	687	.53	.617	.54	.623	.58	.647
VH581463	1,383	350	.55	.629	.57	.597	.56	.643
VH581460	1,468	341	.55	.653	.56	.657	.59	.639
VH337306	3,720	908	.62	.658	.63	.670	.64	.673
Average			*.60*	*.655*	*.61*	*.656*	*.62*	*.661*

Note. QWK = quadratic weighted kappa.

average adjusted R^2 of .61 and achieved an average quadratic weighted kappa of .656. The trait-based model achieved an average adjusted R^2 of .63 and an average quadratic weighted kappa on the test set of .661. Once again, the trait-based model showed the highest average performance.

Human-Scored Essays Written to GRE Prompts

A similar pattern of results was observed when we applied the model to predict human scores for GRE essays (see Table 50). Once again, organization had the highest correlation with human scores. It was followed by conventionality, grammar, formality, vocabulary length, vocabulary difficulty, sentence complexity, and lexical tightness.

We did not have enough data in our GRE data set to construct prompt-specific models, but when we constructed task-specific models, the results indicated better performance for the trait model. On argument essays, the trait model has the highest adjusted R^2 and the highest quadratic weighted kappa. Similarly, issue essays demonstrated the strongest performance on the trait model (see Table 51).

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 50 Correlation of Trait Scores to Human Holistic Essay Scores for GRE Prompts

Trait	Correlation
Organization	.74
Formality	.37
Vocabulary length	.32
Vocabulary difficulty	.28
Interactivity	−.14
Sentence length	−.05
Sentence complexity	.27
Conventionality	.50
Grammar	.48
Stance taking	.04
Contextualization	.06
Concreteness	.06
Cohesion	.04
Lexical tightness[a]	.16

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

Table 51 Performance of e-rater and Trait Feature-Based Models for GRE Prompts

			Standard e-rater features		Extended e-rater model		Trait model	
Prompt ID	*N* train	*N* test	Adj. R^2	QWK	Adj. R^2	QWK	Adj. R^2	QWK
Argument	3,882	981	.64	.770	.65	.791	.67	.794
Issue	3,840	950	.62	.748	.63	.749	.66	.769

Note. QWK = quadratic weighted kappa.

Table 52 Correlation of Trait Scores to Human Holistic Essay Scores for TOEFL Independent Prompts

Trait	Correlation
Organization	.68
Formality	.45
Vocabulary length	.41
Vocabulary difficulty	.37
Interactivity	−.18
Sentence length	.05
Sentence complexity	.28
Conventionality	.60
Grammar	.62
Stance taking	−.09
Contextualization	.05
Concreteness	.02
Cohesion	−.04
Lexical tightness[a]	.19

[a] While lexical tightness is not one of the traits in the structural equal model, it is included here because it is orthogonal to those traits and adds variance above and beyond them.

Human-Scored Essays Written to TOEFL Independent Prompts

The results were slightly different for the final set of essays for which we had human scores. In our sample of essays written to TOEFL Independent prompts, the strongest correlation with score was for the organization trait, followed by the grammar trait, the conventionality trait, formality, vocabulary length, vocabulary difficulty, and sentence complexity (see Table 52).

However, as Table 53 demonstrates, the trait model does not outperform the augmented e-rater model on TOEFL Independent essays. Although the trait model had the highest adjusted R^2 in all 10 prompts, it had the best quadratic

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 53 Performance of e-rater Feature- and Trait Feature-Based Models on Human-Scored Essays Written to TOEFL Independent Prompts

			Standard e-rater features		Extended e-rater model		Trait model	
Prompt ID	*N* train	*N* test	Adj. R^2	QWK	Adj. R^2	QWK	Adj. R^2	QWK
VC333628	776	223	.59	.686	.64	.716	.59	.686
VC431370	794	206	.66	.679	.67	.701	.66	.679
VF287760	815	183	.65	.767	.66	.777	.65	.767
VH077294	780	220	.67	.676	.69	.667	.67	.676
VH233786	809	190	.76	.745	.78	.792	.76	.745
VH272550	810	189	.59	.730	.61	.734	.59	.730
VH328318	811	182	.75	.715	.75	.675	.75	.715
VH457434	792	202	.68	.687	.69	.687	.68	.687
VH545052	796	195	.61	.696	.61	.654	.61	.696
VH700281	791	204	.77	.721	.77	.729	.77	.720
Average			*.67*	*.710*	*.68*	*.714*	*.69*	*.713*

Note. QWK = quadratic weighted kappa.

weighted kappa in only half of the models. The model built with standard e-rater features had an average adjusted R^2 of .67 and a quadratic weighted kappa on the test set of .710. The augmented e-rater model had an average adjusted R^2 of .68 and a quadratic weighted kappa on the test set of .714. The trait model had an average adjusted R^2 of .69 but an average quadratic weighted kappa of .713.

Research Question 6c: Discussion

Overall, when we used the trait scores to build an essay scoring model, it performed better than a model using standard e-rater features. The trait model also performed slightly better than an e-rater model supplemented by features that are not part of the standard e-rater model but are incorporated into the trait-based model (e.g., grammaticality and lexical tightness). The trait-based model had superior performance on every data set, except for TOEFL Independent prompts, for which its performance was roughly comparable to that of the augmented e-rater model.

In one sense, these finding are not surprising, because the trait-based model includes most (though not all) of the features e-rater uses to predict scores, and these features were selected precisely for their score prediction capacity and stability across tasks. However, the trait model combined e-rater features with other features (and with one another), reducing the number of parameters that could be set for these features to maximize score prediction, and it introduced other features that were more strongly connected with genre differences than with score.

More generally, the inclusion of the genre-related traits appeared to contribute to the model's ability to predict scores accurately. If we applied a stepwise regression technique to our various data sets, instead of entering all variables simultaneously, the weaker, genre-related traits tended to be excluded from the models, but both adjusted R^2 and quadratic weighted kappa on the test set decreased consistently across models and prompts. And in the case of the CBAL data, for which we were able to test genre-specific models, the genre-differentiating features had patterns of significant weights that varied depending on the genre of the model.

These results are consistent with the fact that in the Criterion operational data, we observed significant trait differences by genre across all of the traits. Essay quality may well suffer if students write essays that deviate significantly from the genre norms appropriate to the writing task. These genre effects may not have a large impact on score, unless the deviation from genre norms is also large. But the stronger predictive performance of the trait model suggests that including genre norms in an essay scoring model is likely to improve its overall performance.

Overall Discussion and Conclusions

The purpose of this study was to explore approaches that would model dimensions of variation in student texts using automated features that were extracted using NLP techniques. If we can identify and measure major dimensions of textual

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

variation, then we can use them to assess student performance, track changes in performance over time, and provide appropriate reporting and feedback, either for students or for teachers.

Our model draws on feature sets that were devised for different purposes—to measure writing quality, to assess text readability, and to measure register and genre differences among texts. Therefore our model intends to provide a broad and flexible range of metrics that can be used to characterize differences among student texts on a variety of dimensions. We do not claim that the feature set we use is exhaustive in this regard. For instance, features could be usefully added to capture characteristic features of descriptive, causal, analytical, or reflective writing, which we have not addressed here (Klebanov & Madnani, 2022). But because the model incorporates features that have been validated for a variety of purposes, and that have been validated with a broad range of writing tasks and texts, we can have reasonable confidence in its utility and significance.

Overall, the theoretical model we have proposed fits our classroom K–12 data acceptably well and can be extended without significant loss of modeling accuracy to a variety of (mostly essay) writing tasks ranging from K–12 to college. The model does not fit longer responses (such as college research papers) particularly well, but overall, the trait model appears to provide a reasonably accurate description of important dimensions of variation in student writing.

The model contains some traits that are strongly related to score and less strongly related to genre, such as organization, formality, lexical tightness, grammar, and conventionality. It contains other traits that are only weakly related to score but are strongly related to genre, such as contextualization, stance taking, concreteness, and cohesion. However, the genre-differentiating traits also contribute to score prediction, and the score-differentiating traits also help to differentiate among genres. As a result, the combined model makes it possible to describe performance patterns both within and across genres. In principle, we could use trait patterns to describe differences between groups of students, though in this study, we found only small demographic differences.

The trait model appears to achieve sufficient reliability for the traits to be used as subscores, at least with respect to the traits that are more strongly associated with score differences. Perhaps most critically, our results indicate that a student's trait profile on one essay provides better prediction of their performance on the same task, as well as future tasks, than either human or e-rater holistic scores alone. Examination of longitudinal data also indicates that the traits with the strongest associations with score demonstrate patterns of growth, both within and between grades and after revision for the same assignment. This provides strong validity evidence favoring the use of the trait scores to describe student writing performance.

Overall, the results we report here indicate that the trait model is valid and reliable for assessing and describing student writing performance.

Notes

1 Available from https://github.com/EducationalTestingService/ies-writing-achievement-study-data
2 Available from https://lsa.umich.edu/eli/language-resources/micase-micusp.html
3 When teachers or administrators create an assignment in Criterion, they control various options, including the grade level assigned to the class; the genre (expository or persuasive) of the task; the number of revisions allowed; whether work must be completed by a deadline or within a time limit; whether the task will be associated with peer groups; and the availability of various writing tools, such as spell-check and thesaurus. These features are relevant to the conditions under which students submitted their work and were therefore recorded as part of the underlying data set, though they did not appear to make a significant difference in the statistical properties of the submitted essays and are therefore excluded from the results described in this report.
4 Some of the loadings are negative where we would expect positive associations, due to collinearity between superordinate and subordinate traits.
5 We excluded formality as it is strongly associated with vocabulary length and vocabulary difficulty, the features of which cross-load on formality. If we include formality, whether or not we also include vocabulary length and vocabulary difficulty, formality is not significant, and the variable inflation factor (VIF) for these variables is very high, which is why we present the regression shown in Table 16.
6 We excluded vocabulary length and vocabulary difficulty, as they were strongly associated with formality. Formality is significant whether or not these features are included, but if they are included, the VIF is very high, indicating distortion of the weights due to collinearity.
7 Note that the genre means we report are by definition means for Criterion topic library prompts, which were designed as stand-alone writing assignments that did not require students to address specific academic content. As a result, some of the

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

means are skewed relative to the sample as a whole, which included teacher-created, source-based writing assignments. For instance, the Criterion topic library prompts appear to be lower on formality (academic language) than the entire pool of essay submissions, presumably because teacher-created assignments are more likely to target specific academic content.

8 We could not include both superordinate and subordinate traits in the regressions in these SEM models without causing overfitting due to collinearity, which was important in this context because we wanted accurate estimates of the effects of each writing trait.

9 We do not report the pattern of individual feature weights here because we used ordinary least squares regression, and the superordinate and subordinate traits are necessarily collinear. Although the use of collinear features does not affect the score prediction results, it does lead to inflated individual weights. This is not a problem in the present context, where we are concerned primarily with the predictive value of the trait model, but for other purposes, such as explaining the relative contribution of the features in the trait model, some features would have to be excluded or else a more sophisticated modeling method would be necessary.

10 Results for these models were as follows: literary analysis, policy argument, and cost–benefit analysis.

References

Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, *108*(7), 911–924. https://doi.org/10.1037/edu0000109

Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (Research Report No. RR-11-39). ETS. https://doi.org/10.1002/j.2333-8504.2011.tb02275.x

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (Research Report No. RR-08-19). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02105.x

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for TOEFL Writing tasks* (Research Report No. RR-15-14). ETS. https://doi.org/10.1002/ets2.12061

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K–12 assessments* (Research Report No. RR-11-23). ETS. https://doi.org/10.1002/j.2333-8504.2011.tb02259.x

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–44. https://doi.org/10.1515/ling.1989.27.1.3

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *tesol Quarterly*, *36*(1), 9–48. https://doi.org/10.2307/3588359

Brinton, J. E., & Danielson, W. A. (1958). A factor analysis of language elements affecting readability. *Journalism Quarterly*, *35*(4), 420–426. https://doi.org/10.1177/107769905803500402

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25*(3), 27.

Burstein, J., Elliot, N., Beigman Klebanov, B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, *2*(1), 258–313. https://doi.org/10.37514/JWA-J.2018.2.1.12

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2019, April 4–8). *Linking writing analytics and broader cognitive and interpersonal outcomes* [Paper presentation]. National Conference on Measurement in Education, Toronto, ON, Canada.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., Ling, G., & Holtzman, S. (2019). Exploring writing analytics and postsecondary success indicators. In J. Cunningham, N. Hoover, S. Hsiao, G. Lynch, K. McCarthy, C. Brooks, R. Ferguson, & U. Hoppe (Eds.), *Companion proceedings of the 9th international conference on Learning Analytics and Knowledge (LAK'19)* (pp. 213–214). Society for Learning Analytics Research.

Burstein, J., McCaffrey, D., Klebanov, B. B., & Ling, G. (2017, September 8). *Exploring relationships between writing and broader outcomes with automated writing evaluation* [Paper presentation]. 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 267–281). Taylor and Francis.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Taylor and Francis.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Culham, R. (2003). 6 + 1 traits of writing: The complete guide Grades 3 and up. *Scholastic*.
Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002
Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, *10*(3), 257–275. https://doi.org/10.1207/s1532799xssr1003_4
Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R., Sabatini, J., & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing*, *32*(6), 1575–1606. https://doi.org/10.1007/s11145-018-9852-7
Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. RB-61-15). ETS. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x
Entin, E. B., & Klare, G. R. (1978). Factor analyses of three correlation matrices of readability variables. *Journal of Reading Behavior*, *10*(3), 279–290. https://doi.org/10.1080/10862967809547279
Foltz, P., & Rosenstein, M. (2019). Data-mining large-scale formative writing. In C. Lang, G. Siemens, A. Wise, & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 199–210). Society for Learning Analytics Research. https://doi.org/10.18608/hla17.017
Frow, J. (2014). *Genre* (2nd ed.). Routledge. https://doi.org/10.4324/9781315777351
Gallagher, K. (2011). *Write like this: Teaching real-world writing through modeling and mentor texts*. Stenhouse.
Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636
Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17*(3), 307–352. https://doi.org/10.1177/0741088300017003001
Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424
Kellogg, R. T. (2001). Competition for working memory among writing processes. *American Journal of Psychology*, *114*(2), 170–191. https://doi.org/10.2307/1423513
Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, *37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002
Klebanov, B. B., & Madnani, N. (2022). Genre- and task-specific features. In B. Beigman Klebanov & N. Madnani (Eds.), *Automated essay scoring: Synthesis lectures on human language technologies* (pp. 101–153). Springer. https://doi.org/10.1007/978-3-031-02182-4_6
Madnani, N., Burstein, J., Elliot, N., Klebanov, B. B., Napolitano, D., Andreyev, S., & Schwartz, M. (2018, August). *Writing Mentor: Self-regulated writing feedback for struggling writers* [Paper presentation]. 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, NM, United States.
McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Pschology Review*, *8*, 299–325. https://doi.org/10.1007/BF01464076
McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1
National Center for Education Statistics. (2003–2018a). *Private School Universe Survey (PSS)*. U.S. Department of Education. https://nces.ed.gov/surveys/pss/
National Center for Education Statistics. (2003–2018b). *Public Elementary/Secondary School Universe Survey*. U.S. Department of Education. https://nces.ed.gov/ccd/pubschuniv.asp
Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater® scoring engine* (Research Report No. RR-09-01). ETS. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x
Ramineni, C., & Deane, P. (2016). The Criterion® online writing evaluation service. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 163–184). Taylor and Francis. https://doi.org/10.4324/9781315647500-12
Reppen, R. (2007). First language and second language writing development of elementary students. In Y. Kawaguchi & T. Takagaki (Eds.), *Corpus-based perspectives in linguistics* (pp. 147–167). John Benjamins. https://doi.org/10.1075/ubli.6.12rep
Römer, U., & Swales, J. M. (2010). The Michigan Corpus of Upper-Level Student Papers (MICUSP). *Journal of English for Academic Purposes*, *9*(3), 249. https://doi.org/10.1016/j.jeap.2010.04.002
Sheehan, K. M. (2013). Measuring cohesion: An approach that accounts for differences in the degree of integration challenge presented by different types of sentences. *Educational Measurement: Issues and Practice*, *32*(4), 28–37. https://doi.org/10.1111/emip.12017
Sheehan, K. M. (2016). *A review of evidence presented in support of three key claims in the validity argument for the TextEvaluator text analysis tool* (Research Report No. RR-16-12). ETS. https://doi.org/10.1002/ets2.12100
Sheehan, K. M., Kostin, I., & Futagi, Y. (2007, August 1–4). *Reading level assessment for literary and expository texts* [Paper presentation]. Annual meeting of the Cognitive Science Society, Nashville, TN, United States.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, *115*(2), 184–209. https://doi.org/10.1086/678294

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62*(1), 5–18. https://doi.org/10.1177/0013164402062001001

van Rijn, P., Chen, J., & Yan-Koo, Y. (2016). *Statistical results from the 2013 CBAL English language arts multistate study: Parallel forms for policy recommendation writing* (Research Memorandum No. RM-16-01). ETS. https://www.ets.org/research/policy_research_reports/publications/report/2016/jvwn.html

van Rijn, P., & Yuen, Y.-K. (2016). *Statistical results from the 2013 CBAL English language arts multistate study: Parallel forms for argumentative writing* (Research Memorandum RM-16-15). ETS. https://www.ets.org/research/policy_research_reports/publications/report/2016/jwxy.html

Weigle, S. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Suggested citation:

Deane, P., Yan, D., Castellano, K., Attali, Y., Lamar, M., Zhang, M., Blood, I., Bruno, J. V., Li, C., Cui, W., Ruan, C., Appel, C., James, K., Long, R., & Qureshi, F. (2024). *Modeling writing traits in a formative essay corpus* (Research Report No. RR-24-02). ETS. https://doi.org/10.1002/ets2.12377

Action Editor: Klaus Zechner

Reviewers: Matt Johnson and Omid Kashef

CRITERION, E-RATER, ETS, the ETS logo, GRE, TEXTEVALUATOR, TOEFL, and WRITING MENTOR are registered trademarks of Educational Testing Service (ETS). CBAL is a trademark of ETS. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database.

23308516, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12377, Wiley Online Library on [30/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License